USING SIGE HBTS FOR QUANTUM SCIENCE AT DEEP CRYOGENIC TEMPERATURES

A Dissertation
Presented to
The Academic Faculty

By

Hanbin Ying

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2021

USING SIGE HBTS FOR QUANTUM SCIENCE AT DEEP CRYOGENIC TEMPERATURES

Thesis committee:

Dr. John D. Cressler, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Dragomir Davidović
School of Physics
Georgia Institute of Technology

Dr. Asif Islam Khan
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Martin Mourigal
School of Physics
Georgia Institute of Technology

Dr. Nelson E. Lourenco
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date approved: Feb 15, 2021

To my parents, who unequivocally support me.

To my wife, who stays with me through all the ups and downs in my life.

ACKNOWLEDGMENTS

Along the journey of this research, so many people gave me big help and endured me with patience. Without them, this work will not be possible. Although there is simply more gratitude than I can include in a few words here, I would still like to at least scratch the surface and extend my sincere thanks to a few people.

First, I would like to thank my family. I want to thank my parents, who gave me more than enough (and sometimes too much) freedom to make decisions for myself and bear all consequences as a result of my decisions. The lessons I learned throughout my life journey so far give me the confidence to stick my neck into the prestigious Ph.D. program at Georgia Tech. I also want to thank my wife, who is an integral part of my family and my life, and has stayed with me for the past eleven years (and still counting). Her faith in me never faded even in the five-year long-distance relationship that has a 12-hour time difference. Her confidence in my ability and my future gave me the strength to get through the frustration and anxiety of the Ph.D. program and persevere through any hardship. Without my family, I would probably never have entered or finished the Ph.D. program.

Next, I would like to sincerely thank my advisor, Dr. Cressler, whose decision to take me in the team in 2015 completely changed my life trajectory. His extensive technical knowledge and mentoring has nurtured my growth in all these years and allow me to become a better thinker, engineer, and designer. His wisdom brings out my potential and makes me a better person not only in research but in all facets of life. His frequent reminders of work-life balance teach me to take care of my life outside research. As a result, I get to see that life is made of many exciting aspects in addition to the research. As a philosopher, Dr. Cressler inspires me to appreciate all the nice things in this world.

I would also like to thank Dr. Khan, Dr. Davidovic, Dr. Mourigal, and Dr. Lourenco, who are willing to spare their valuable time, serve on my Ph.D. dissertation committee, and give me a lot of valuable feedback on my research as I get along the progress.

I want to express my gratitude to my colleagues including Luwei and Jason from the Physics Department, and Jeff, Sunil, Yunyi, Milad, Brian, Uppili, Anup, Nelson, Adrian, George, and Zach from the SiGe Team. All of them, at one point or another, spent a tremendous amount of time teaching me valuable skills or concepts, such as in experiment preparation, device physics, and circuit theory. At other times, they simply have the patience to check grammars in my writing, debug TCAD or Cadence problems with me, and answer my long questions. Without them, I would have been tripped over by many seemingly insurmountable hurdles and the progress towards my Ph.D. would have been much slower. I would also like to thank everyone else in the SiGe Device and Circuit Team. It is simply a wonder how much I learned from these people in casual conversations.

I'm indebted to the administrative staff members at TSRB and ECE for their support, including but not limited to Alison, Scott, and Carolyn, without whom I will not be able to purchase the right tools and carry out the research plans on time.

I would like to formally thank my sponsors at Sandia Corporation, including T. England, M. Holmes, and S. Carr, and sponsors at Fermilab and Jet Propulsion Lab (JPL), including D. Braga, F. Fahim, M. Shaw, and B. Korzh. With their financial sponsorship and technical guidance, I can push through the boundary of science and engineering.

Last but not least, I would like to thank A. Joseph, V. Jain, N. Cahoon, T. Lamothe, D. Harame and the GlobalFoundries SiGe team for their support on wafer fabrications.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

BiCMOS technology can provide tremendous benefits for quantum science applications at deep cryogenic temperatures as low as milliKelvin. In particular, cryogenically operated SiGe HBTs have proven low-noise, low-power, and high-bandwidth performance, all of which is critically needed for high-fidelity qubit readout in quantum computing and low-jitter time-tagging of superconducting nanowire single-photon detectors (SNSPD). The CMOS compatibility inherent to BiCMOS technology enables dense digital circuits for qubit control in quantum computing or post-processing of photon arrival time. Overall, BiCMOS technology operated at cryogenic temperatures can satisfy the need for the scaling of future large-scale quantum computers, SNSPD array, and other quantum science applications.

The objective of this research is to investigate the feasibility of using BiCMOS technology for these quantum science applications and clear some major roadblocks. The requirement for these applications are detailed, and the research is conducted in a systematic way targeting four important aspects of SiGe HBTs, namely, cryogenic characterizations, device physics, compact modeling, and circuit designs.

The following is a summary of the contributions of this work:

1. The characterization of SiGe HBTs at a record low temperature of 70 mK. This characterization was published in *IEEE Electron Device Letters* in 2016 [1].

2. A method to differentiate the transport mechanism of collector current at cryogenic temperature, and a study into the effect of future technology scaling on the transport mechanism. This method and study were published in *IEEE Transactions on Electron Devices* in 2018 [2].

3. Characterizations of the DC and RF parameter variability of SiGe HBTs operating at cryogenic temperatures. This work was presented in *2019 IEEE BiCMOS and Com-*

pound Semiconductor Integrated Circuits and Technology Symposium (BCICTS) [3].

4. An investigation into the physical origin of the variability in SiGe HBTs and PN junctions, and possible mitigation methods. This work will be published in *IEEE Transactions on Electron Devices* in 2021 [4].

5. Construction of a compact model of SiGe HBTs operating at cryogenic temperatures, verified by both device and circuit-level correlation. This work was presented in *2020 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS)* [5].

6. A study of the effect of device variability at the circuit level, including its effect on current mirrors and low noise amplifiers. This work will be published in *IEEE Transactions on Electron Devices* in 2021 [6].

Copyright Statement

CHAPTER 1
INTRODUCTION

Quantum information science is an emerging and rapidly evolving field that is a multi-discipline synthesis of three important scientific fields in the 20th century, namely, quantum physics, information theory, and computer science [7].

In the 21st century, the scientific advancement and nascent commercialization of quantum computers (QC) come into the central role in quantum information science. Unlike classical computers, quantum computers use quantum bits (qubits) to process information [8]. In quantum computers, certain algorithms can run much faster than in a classical supercomputer. Such algorithms include Shor's algorithm for integer factorization, Grover's algorithm for search and optimization, and quantum simulation algorithms for simulating many-particle systems [9]. The speed advantage of these algorithms compared to classical computers can bring new breakthrough in secure quantum network, accurate molecule models in simulations, optimization with large and unstructured datasets, faster runtime for artificial intelligent algorithms, etc, all of which have a far-reaching impact on cryptography, chemistry, finance, artificial intelligence, etc[10].

However, quantum computing is still in infancy in regard to the promises above. A significant breakthrough in both physics and engineering is needed to bring its full power to life. For example, qubits are simultaneously powerful and fragile. To ensure the correctness of computations, quantum error-correction (QEC) is needed, which requires many physical qubits to construct one logical qubit [11]. To achieve the performance useful for practical applications, it is estimated that millions of physical qubits are needed to construct thousands of logical qubits [12, 13]. Since each qubit needs to have a number of accompanying electronics for proper functions (e.g. initialization, control, interaction, readout), the need for more qubits translates to the need for more electronics, and more importantly, the

need for electronics that are scalable. When it comes to the scaling of man-made objects, the crowning example is the scaling of transistors in integrated circuits. As predicted by the Moore's Law, the number of transistors per silicon chip has been doubling every two years, which leads to many consequential technologies such as phones, cameras, televisions, and other smart devices. If quantum computers can scale at such speed, it will open a new era of technological prosperity for mankind. The thirsty search and development for scalable quantum computers is the high-level motivation for the present thesis.

The extensive activities and breakthroughs in quantum computing also incubate large development in many neighboring quantum science fields. For example, high-performance quantum detectors have seen a large improvement in recent years. In particular, superconducting nanowire single-photon detectors (SNSPDs) emerges with a record-breaking performance in sensitivity, speed, jitter, and efficiency [14]. As a result, SNSPDs have emerged in late-stage research and commercialization in fields like quantum key distribution, deep-space communication, commercial LiDARs, and optical quantum computers [15].

Quantum computers and SNSPDs are merely two representative applications in the realm of quantum science. Many exciting physics await the discovery and improvement that is enabled and enhanced by having access to integrated electronics. In particular, as will be shown in the next section, these applications and experiments can greatly benefit from integrated electronic circuits operating at ultra-low temperatures (ranging from milliKelvin to a few Kelvin).

The remaining portion of this chapter will explain quantum computing and SNSPD with sufficient details to provide the understanding and appreciation of why cryogenic integrated electronics are needed in quantum science. Subsequently, silicon-germanium heterojunction bipolar transistors (SiGe HBTs) and BiCMOS technology are shown to be a promising candidates for realizing those electronics but require further research in some key areas, some of which are tackled in the rest of this thesis.

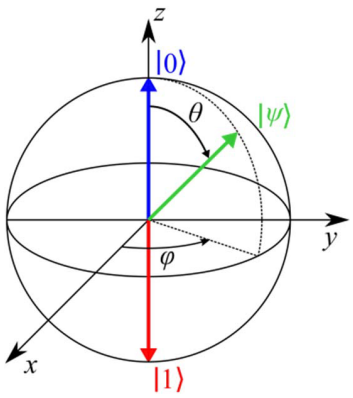

Figure 1.1: The Bloch sphere representation of quantum bits. (After [18])

1.1 Quantum Computing

Due to the extensive activities, an abundance of textbooks and publications exist that explain virtually all aspects of quantum computing [8, 16]. Here, we merely reproduce some key facts that are relevant for the understanding of why electronics are needed. The focus will mostly be on the superconducting qubit or the quantum dot qubit because they are the most compatible with electronics integration.

1.1.1 Operation of Quantum Computers

In quantum computers, the basic element of information is a quantum bit, or qubit. Unlike in classical computers, where each bit of information can take a discrete value of 0 or 1, qubits in quantum computers can take a value of 0, 1, or a superposition of 0 and 1 [17]. This is best visualized in the Bloch sphere as shown in Figure 1.1. Bloch sphere represents all possible states (including 0, 1, and all superposition states) of a 2-level system (i.e., a qubit). The state 0 or state 1 is traditionally plotted as the north or the south pole of the sphere, while the superposition state is represented as any other points on the surface. The equator of the sphere corresponds to states with an equal probability of being in 0 and 1. Compared to the classical bit, which can only be either 0 or 1 at any moment, qubits can theoretically be in a state that is a combination of 0 and 1.

However, if only one qubit is used, the power of the qubit is somewhat limited. No

matter which state the qubit is in, upon measurement by the operator, the qubit will always fall to a deterministic state of 0 or 1. An important concept in quantum computers is the concept of probability amplitude. The probability of measuring the qubit to be in a state of 0 or 1 depends on the squared of the probability amplitude associated with each state. The amplitude can be negative or even complex numbers. With only one qubit, the negative or complex value in the probability amplitude will not make any difference from an experimental point of view, because the squared of these numbers is always positive.

To harness the full power of qubits in quantum computing, quantum entanglement is needed, where two or more qubits are entangled. Entanglement allows the probability amplitudes of multiple qubits to interfere (e.g., add, subtract, etc) with each other. The interference of multiple qubits is very similar to the interference of lights in optics, where constructive interference gives a maximum while destructive interference gives a minimum of probability amplitude. Once the amplitude is squared after the interference, the probability of measuring the qubits in a particular state will also have a minimum and maximum depending on the interference. Quantum computers, in essence, manipulate the environment and the interactions between qubits in such a way that the answer to the problem will come out with a high probability, while the incorrect answer has a low probability. Then, when we read out the qubit, the answer is encoded in the probability of obtaining a particular measured state (0 versus 1). Compared to classical bits, where N bits can only go through 2^N possible values *sequentially* (i.e., one by one, but not at the same time), N qubits in quantum computing can go through 2^N values *simultaneously* because the probability amplitude of every qubit is interacting with that of other qubits simultaneously. If designed properly, a quantum algorithm can access the 2^N values simultaneously and find the answer from those values quickly, while classical computers need to go through the values one by one. When $N > 50$, the solution space of 2^N is so large that it takes too large of memory than the universe can hold to solve the problem using classical computers, while it is possible to solve it with quantum computers. The parallelism offered by super-

position and entanglement is the fundamental reason for the exponential speed-up offered by quantum computers.

There are two important notes, however. First, even though the N qubits can access 2^N values during the computation, we can only access the final results, which are just N binary bits for N qubits. In other words, the 2^N states are hidden from us. Second, we can only know the probability of a qubit in a given state by repeating the same computation many times, thus gathering large sets of measurements to claim what the probability is. Therefore, the full solution process of quantum computers is as follows. The qubits are prepared (i.e. initiated to known states) before the computation, left alone during the computation, read out after the computation, and the whole process repeated many times to obtain the probability. Although each computation takes little time (only a few nanoseconds to microseconds, depending on the qubit technology), the total computation time also includes the overhead time of initialization, readout, and the repetition of numerous computations. Therefore, a key challenge is how to perform these actions fast.

Another key challenge in quantum computing is to perform each operation accurately. Due to the extreme fragility of qubits, any interactions with the environment can corrupt the qubit state to a state different from what it is initialized, which results in computation errors. Unlike macroscopic objects in the real world, where we can prevent one object from interacting with another macroscopic object (e.g. place one object in a metallic shield box and the other one outside), in the microscopic world of quantum mechanics, no shield is perfect and everything is interacting or coupled with everything else at all times. The interaction is constantly, albeit slowly (if designed well), changing and corrupting the state of the qubit. In other words, even though the computation is not finished yet, the qubit state is gradually changing from what it is manipulated to be. There is only a finite amount of time to do the computation and readout the result before the qubit state is corrupted.

Given the previous two challenges, a fundamental trade-off exists between good isolation and fast readout. Good isolation means the qubits are isolated from the environment as

well as possible, thus preserving the desirable state and gives the computation the longest duration to complete without corruption. However, perfect isolation also means the experimentalist, being part of the environment from the qubit's perspective, can not obtain useful information from the qubit when the computation finishes, because the readout of a qubit is, effectively, letting the qubit interact strongly with the environment and "leak" information out to the experimentalist. Good isolation gives the computation the best chance to finish, but also slows down the readout, which is problematic because qubits only have a limited lifetime before they are inevitably corrupted by the environment. To make things worse, control and readout functions can introduce additional noise that is leaking into the qubit. The implication for this trade-off is the extremely strict requirement imposed on the surrounding environment of the qubit, including the electronics connected to it.

1.1.2 Electronics for Quantum Computing

Electronics are currently the workhorse for the control and readout of superconducting and quantum dot qubits. An example architecture is shown in Figure 1.2, which shows the two main functionalities, namely, the control and the readout. In superconducting qubit (and various other qubits), control electronics inject RF pulses with a given envelope waveform and frequency to manipulate the qubit to a known state. The waveform is determined by the required manipulation, while the frequency is determined by the particular qubit that the operation is addressing. To visualize, the manipulation comes down to sequences of rotations of qubit state in the Bloch sphere. The rotations can be around the x-axis, y-axis, z-axis, or combinations of them, all of which corresponding to different shapes of the waveform. Therefore, by controlling the pulse to a specific waveform and frequency, we can manipulate a particular qubit to a given state.

Readout electronics operate by injecting RF pulses to a resonator that is coupled dispersively to the qubit, where "dispersive" simply means the resonator's natural frequency is far from the qubit frequency, thus only probing the qubit information slightly (or disper-

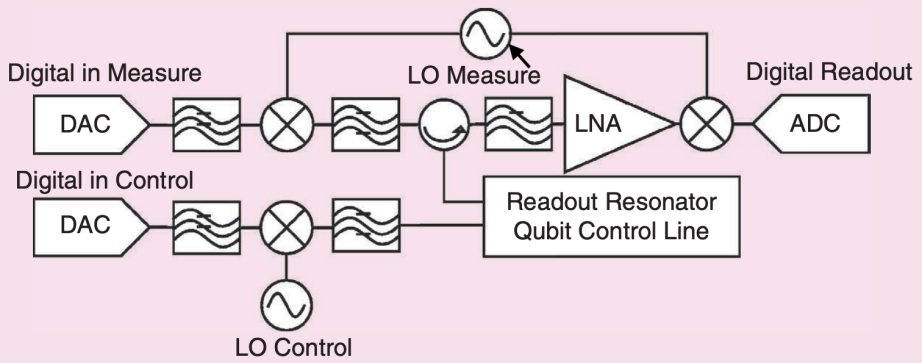

Figure 1.2: An example RF/analog architecture for controlling and reading out qubit. (after [19])

sively). The pulse reflected back from the resonator will be used to determine the qubit state because the impedance of the resonator depends on the qubit state. The frequency of this pulse is typically in the 4–8 GHz for superconducting qubits and up to 20 GHz for quantum dot qubits [20].

As shown in Figure 1.2, each qubit requires a fair amount of control and readout circuits. A critical issue is that all of the control and readout lines need to be RF coaxial cables to support RF pulses. Currently, control and readout functionalities are implemented by equipment placed at room temperature, and the number of cables going into the cryostat is approaching the size and heat load limit, as can be visualized in Figure 1.3. Clearly, having tens of thousands of coaxial cables into a cryostat is not feasible given the cryostat size and heat load constraint. Therefore, ongoing research focuses on integrating some of the functionality into a cryogenic integrated circuit chip, in which case the only wires going to room temperature are the few digital lines for communications. The next section examines what is needed for the cryogenic integrated circuits.

1.1.3 Requirement and Challenges for QC Electronics

As can be seen from Figure 1.2, several functional blocks are needed to control and read out the qubit. On the control side, digital to analog converters (DAC) are required to generate baseband waveforms for a given control function. The waveform is up-converted by the

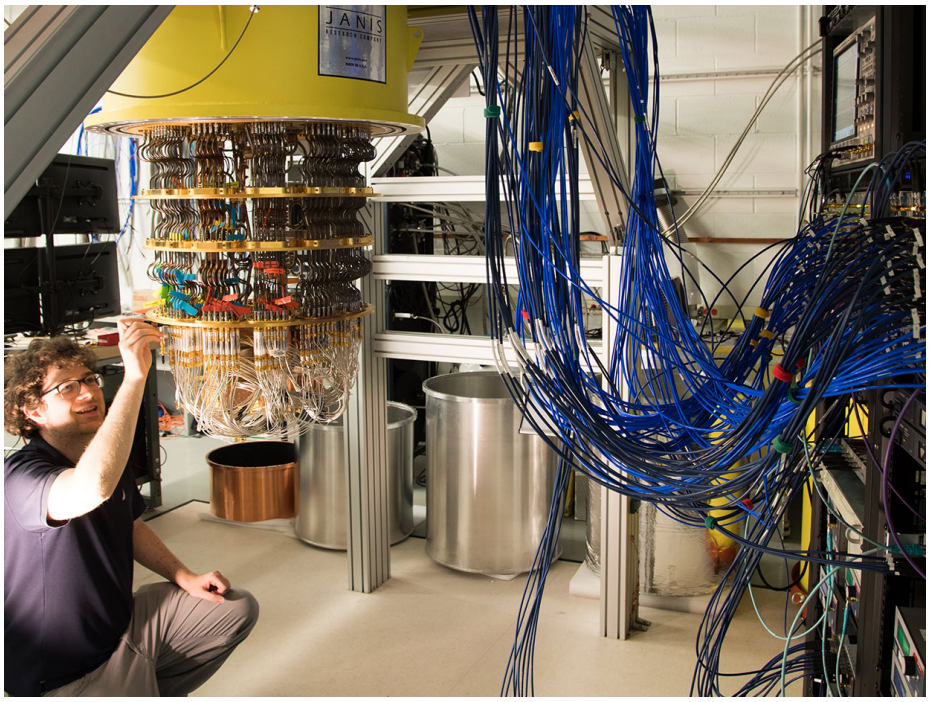

Figure 1.3: An example setup of quantum computers that show the many coaxial cables between the cryostat and room temperature equipment. (after [21])

mixer to the qubit frequency, which is injected into the qubit. On the readout side, another DAC generates a waveform, which is up-converted and injected into the resonator coupled to the qubit. The reflection is amplified by a low noise amplifier (LNA) and down-converted to the baseband, which is then digitized with analog to digital converters (DACs) for further processing. As can be seen, multiple circuit blocks are needed for each qubit. As mentioned before, ideally most of these blocks are situated at cryogenic temperatures to remove the costly and bulky RF cables going into the cryostat.

However, there are several challenges for placing such electronics in the cryostat. First, the frequency of the local oscillator (LO) signals that supplies the mixer needs to be stable and have low phase noise, because phase noise is directly injected into the qubit and causes degradation of qubit states. It also needs to have low spectral leakage because another qubit is operating at a nearby frequency. Such requirements are not difficult at room temperature, where signals are generated by voltage-controlled oscillators (VCO) and locked to a clean frequency source such as quartz crystal oscillator by a phase-locked loop (PLL). At

low temperatures, however, a new system architecture is required, and power consumption is severely limited by the cryostat's cooling power. The mixers used for up and down-conversion also needs to have a low noise figure to prevent excessive noise directly into the qubit. Typically, low noise translates to a decent amount of power consumption, which requires the electronics to be far away from the qubit to prevent too much heat dissipation. However, because the wavelength at 8 GHz is only 37 cm (and even shorter at a higher frequency), the pulse signals cannot go through long cables without attenuation and dispersion (due to transmission line effect), which affects the pulse shape and causes qubit errors if not calibrated. Therefore, the electronics are best placed closer to the qubit, while it creates a more stringent power consumption limit allowed by the cryostat.

On the readout side, the LNA needs to have low noise because the output to be amplified tends to be extremely small (of the order of -125 dBm) [17]. Typically, this requires an amplifier noise temperature of merely 1-2 K, which is theoretically impossible if the amplifier is situated at room temperature. This further demonstrates the need for amplifiers operating at cryogenic temperatures.

Third, since thousands to millions of qubits are needed for an error-corrected quantum computer, the total number of circuits is likely large. In the meantime, the total cooling power of the cryostat can not increase exponentially. This creates a strict limitation on the total power consumption for electronics. As a result, devices used in the circuits will need to operate at low to medium current density, where the RF performance may not be optimal.

Overall, various electrical circuit blocks are needed for quantum computers and they create a unique application requirement for the underlying technology.

1.2 Superconducting Nanowire Single-photon Detectors (SNSPD)

SNSPD use a novel idea to harness the power of superconductors. It is known for a long time that when a superconductor is heated above the superconducting temperature (critical

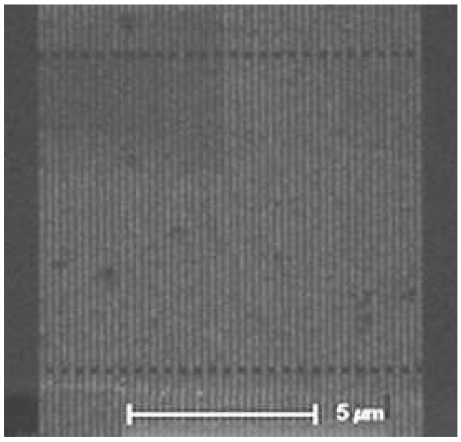

Figure 1.4: An example layout of SNSPD. (after [15])

temperature), it stops behaving as the superconductor and has a finite resistance. It is also known that when a photon hits a material, the photon can be absorbed by the molecules or atoms of the material if the photon provides the bandgap energy to promote the valence electrons. SNSPD combines these two facts to make a versatile photon detector that changes resistance when a single photon hits.

1.2.1 Operation of SNSPDs

SNSPDs are typically constructed as long meandering lines made of a superconductor, as shown in Figure 1.4. At the beginning of a detecting cycle, the SNSPD is biased with a small current (typically a few μA), which is small enough that it does not heat up the superconductor above critical temperature or exceed the critical superconducting current, thus keeping the SNSPD in a superconducting state with zero voltage drop across it. When a photon hits any part of the meandering line, the material absorbs the photon, whose energy is transferred to the superconductor lattice at that absorption spot (called "hot spot"). This creates enough heat and quenches the superconductivity locally, thus giving the SNSPD a finite resistance. Since it is biased with the same current, the sudden appearance of resistance generates a voltage pulse in the time domain that corresponds to a photon hit, similar to what is shown in Figure 1.5(c). After some time, the heat dissipates away from

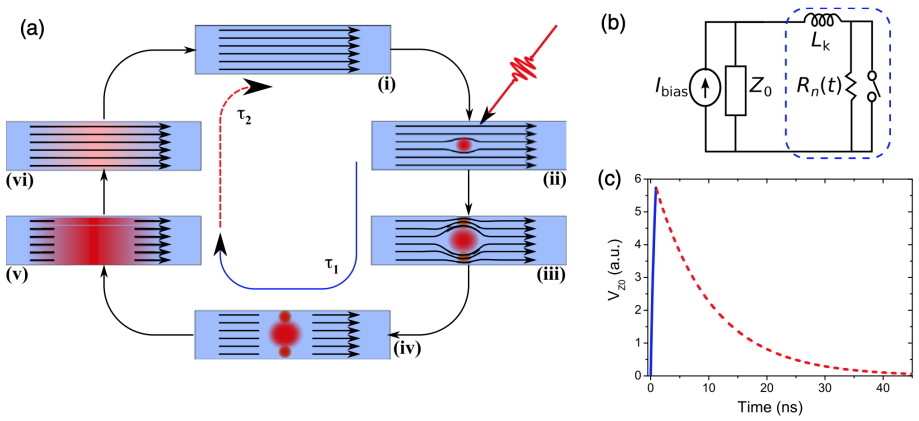

Figure 1.5: (a) The illustration of photon hitting an SNSPD. (b) Simple equivalent model of an SNSPD. (c) Approximated output voltage pulse of the SNSPD when a photon hits. (after [15])

the region, and the superconductor gradually recovers back to its superconducting state, ready for "firing" when the next photon hits. This full cycle is depicted in Figure 1.5(a). In the electrical domain, SNSPDs have been modeled with the equivalent circuit shown in Figure 1.5(b), with a switchable resistance and some inductance. The inductance is to model the transmission line effect on the voltage pulse from the long meander line.

Compared to other photon detectors, SNSPDs have multiple advantages [22]. Different from conventional single-photon avalanche diodes (SPADs), which also have single-photon sensitivity, SNSPDs have a faster count rate (typically $>$ 10 MHz compared to 2-10 MHz for SPADs) due to smaller reset time (5-10 s compared to 50-100 s for SPADs). SNSPDs also have less timing jitter than other detectors because the rising edge is very sharp. Typical jitter from SNSPD is tens of picoseconds, while it is hundreds of picoseconds for SPADs [22]. The state-of-the-art jitter from SNSPD is less than 3 picoseconds [23]. Due to these advantages, the SNSPD is a promising candidate for quantum key distribution (needs low jitter), deep-space optical communication (needs high photon number resolution), and kilometer-scale laser ranging (needs low jitter and high resolution).

One drawback of the SNSPD is the requirement of cryogenic operation (at a few K) to maintain the superconductivity. This limits the placement of detection systems to within the physical boundary of the cryostat, and the cabling between cryostat and room temper-

ature needs to be minimized to reduce heat load, similar to the issues encountered in QC mentioned before.

1.2.2 Requirement and Challenges of Electronics for SNSPDs

SNSPDs require high-speed amplifiers to amplify the voltage pulses and the lower end of the amplifier bandwidth needs to cover a few kHz (near 100 kHz) to preserve the frequency content [24]. For some applications, such as data communication, it is desirable to maintain the DC level in the output waveform, in which case the lower bandwidth needs to extend to DC.

There are two distinct application targets for SNSPDs that have different requirements for the electronics interfacing with SNSPDs. One application requires low jitter for high temporal (time-domain) resolution. The other application requires large arrays of SNSPD for high spatial resolution. For the rest of the applications, the requirement tends to fall somewhere in between these two extremes.

Low jitter is needed in femtosecond photon detection, quantum key distributions, high-speed quantum photonic circuits, and high-resolution laser ranging [25, 26, 27]. The state-of-the-art jitter performance of SNSPD is only a few picoseconds [23]. Therefore, the amplifier needs to introduce a comparable or smaller jitter.

Jitter corresponds to fluctuation in time between the photon incident on the detector and the detector output registering an electrical signal. Based on [28], jitter can come from the intrinsic process of SNSPDs that is of a random nature [29], the geometric structure of SNSPDs [30], and the noise introduced by the readout amplifier [28]. Among this, the geometric jitter was found to be minimized if a differential amplifier is used [31]. For the amplifier jitter, it is affected by amplifier gain, bandwidth, and noise figure, among other factors [28]. A high gain (> 20 dB), large bandwidth (> 3 GHz), and low noise figure or noise temperature (< 10 K) are needed to achieve a low jitter of around a few picoseconds.

Another camp of applications for SNSPDs, such as dark matter detection and space-

based astronomy, require array operation to achieve high pixel resolution (similar to the sensors in camera) [32, 33]. For such applications, the main goal is to have as many SNSPDs in the array as possible for the highest spatial resolution, while the jitter is not necessarily of paramount importance. As a result, low cost (i.e., chip area) and low power amplifiers are critical, because many amplifiers are needed and the total cooling power budget of the cryostat can not be exceeded to preserve the cryogenic environment needed for SNSPDs operation. In addition, a fair amount of digital processing circuits are needed to process the information from each pixel (e.g., time-tagging) before transmitting the information to room temperature computers. The same reasoning as in quantum computing applies here, that only a limited number of cables can be accommodated to go into the cryostat due to heat load and size constraints.

In addition, the voltage pulse from SNSPDs has a broad frequency content from tens of kHz to a few GHz. Similar to quantum computing, the transmission line effect will start to affect the high-frequency portion of the waveform when the cable length is comparable to the wavelength. Since the cable from the SNSPD to outside the cryostat is typically on the order of meters, the dispersion introduced by the long cable requires an elaborate calibration routine. To avoid this, cryogenic amplifiers can be placed near the SNSPDs to reduce the cable length. In addition to the dispersion effect, degraded signal to noise ratio (SNR) is another reason to avoid long cables, especially since the voltage pulse of SNSPD is only of the order of 1 mV or smaller. Overall, cryogenic amplifiers placed near SNSPDs can greatly improve the SNR, achieve low dispersion of signal and thus low jitter, and reduce the heat load of many cables into the cryostat (though the amplifier's power consumption is critical now).

To summarize, the three main challenges of cryogenic electronics for SNSPD applications are low jitter and low power, and the cryogenic operation, which are the exact same goals pursued by all cryogenic amplifiers, and is consistent with the requirement of LNAs for quantum computing applications.

1.3 SiGe BiCMOS Technology

Given the previous explanations for the need for cryogenic electronics in quantum science applications, this section introduces the basics of SiGe HBTs and BiCMOS technology and explains why BiCMOS technology is a good candidate to serve these applications.

SiGe HBT devices have seen large development since their debut in the 1980s and have since become a strong technology contender in cellular transceivers, television tuners, and high-speed optical transceivers [34]. In these applications, both high-speed transistors and a large number of digital blocks are needed on the same chip. Compared to standard CMOS devices, SiGe HBTs have several advantages, such as higher bandwidth, larger gain, lower noise figure, all at a lower cost than CMOS of comparable performance [35]. Compared to III-IV semiconductor devices, such as InP HBTs, InP high electron mobility transistors (HEMTs), and GaAs HEMTs, which have higher or similar RF performance, SiGe HBTs have the advantage of low-cost integration into standard silicon CMOS process, thus having access to high density and low-cost CMOS devices for large digital circuits.

The typical cross-sectional structure of SiGe HBTs is shown in Figure 1.6. As can be seen, the transport direction of carriers is vertical (from the emitter to the collector), which is in contrast to the standard CMOS whose transport direction is lateral. Since the speed of a device is fundamentally limited by how fast the charge carriers can move across the device, a smaller distance along the transport direction is desired. This translates to a reduced layer thickness for SiGe HBTs and reduced lateral dimension for CMOS. Since the lateral dimension is limited by the feature size defined by the lithography mask, a finer feature size increases the manufacturing cost. In comparison, SiGe HBTs can have smaller layer thickness (controlled by material growth rate) with the same lateral dimension, thus enjoying a higher performance at a comparably lower cost than CMOS.

To understand why SiGe HBTs have good RF performance, we can look at a typical band diagram and the Ge profile of SiGe HBTs along the center cutline of the device as

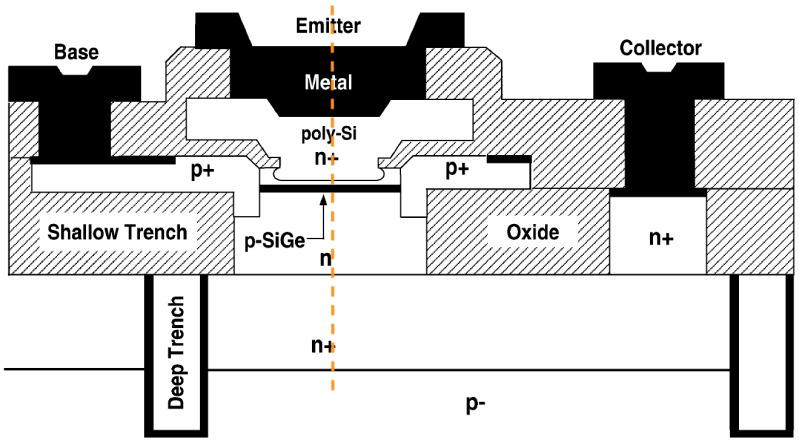

Figure 1.6: Typical structure of SiGe HBTs. The band diagram of the orange cutline (through the center of the device) is shown in Figure 1.7. (after [36])

shown in Figure 1.7. Since Ge has smaller bandgap energy than Si, a larger Ge mole fraction in the SiGe alloy reduces the bandgap, and creates an electric field across the base region. This electric field enables the electrons to move faster from the emitter to the collector, thus giving the transistor a higher speed.

More importantly, the RF performance of SiGe HBTs improves with lower temperatures, as elaborated in [35, 37]. The improvement in current gain is exponential with temperature due to the 1/kT scaling of the Arrhenius function. Figure 1.8 shows the two figure of merits (FoMs) for RF performance, namely, unity current-gain frequency (f_T) and unity power-gain frequency (f_{MAX}) across temperature. As can be seen, at a lower temperature, the speed of the device increases due to lower series resistance and shorter transit time. The DC current gain (β) and the transconductance (g_m) also increases at lower temperatures, as shown in Figure 1.9. The noise figure (NF), which is a critical FoM for LNAs, also decreases at a lower temperature due to the increase in current gain and a decrease in base resistance, as shown in Figure 1.10 [38, 39].

1.3.1 SiGe BiCMOS Technology for Quantum Science

In quantum computing, a large number of computations are needed for accurate calibration and initialization of each qubit. Between computation, the results of the intermediate com-

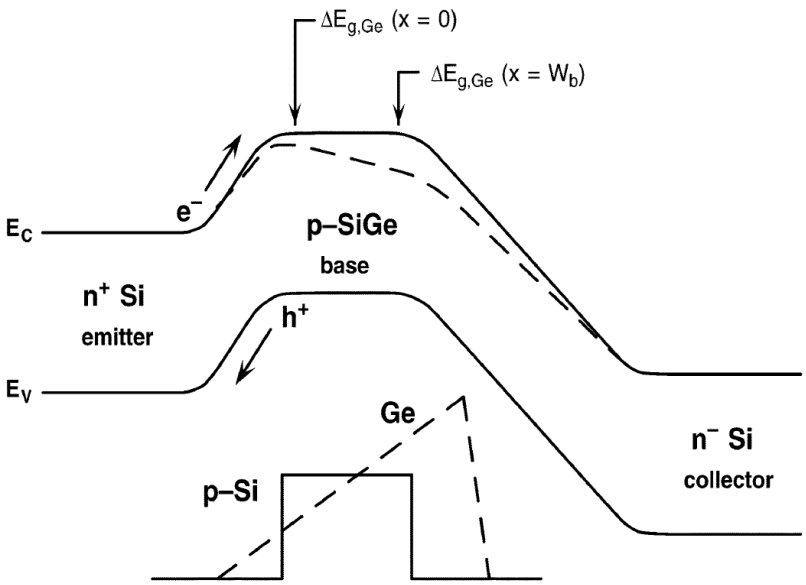

Figure 1.7: Typical band diagram of SiGe HBTs (dashed line) and BJT (solid line). Bottom: an example Ge profile of the SiGe HBT. (after [36])

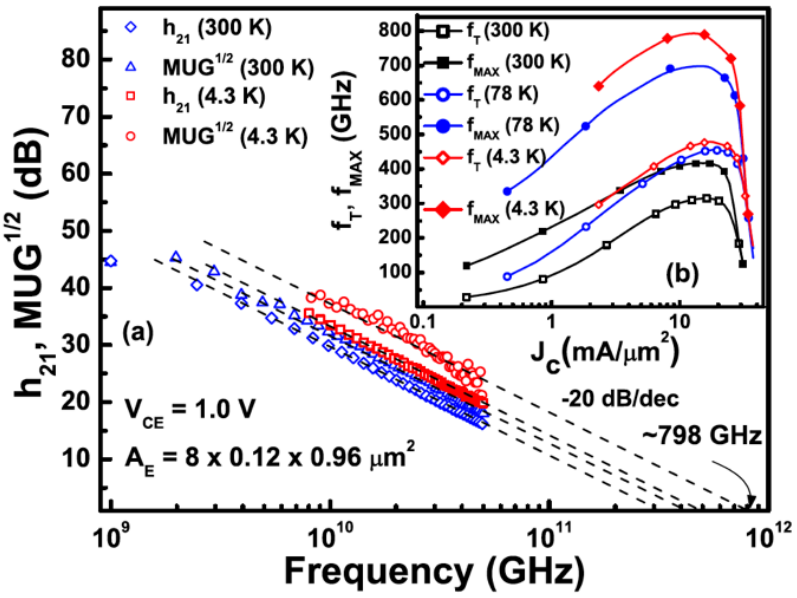

Figure 1.8: The h_{21} parameter of SiGe HBTs across frequency at 300 K and 4.3 K. Inset: f_T/f_{MAX} of SiGe HBTs at 300 K, 78 K, and 4.3 K (after [40]).

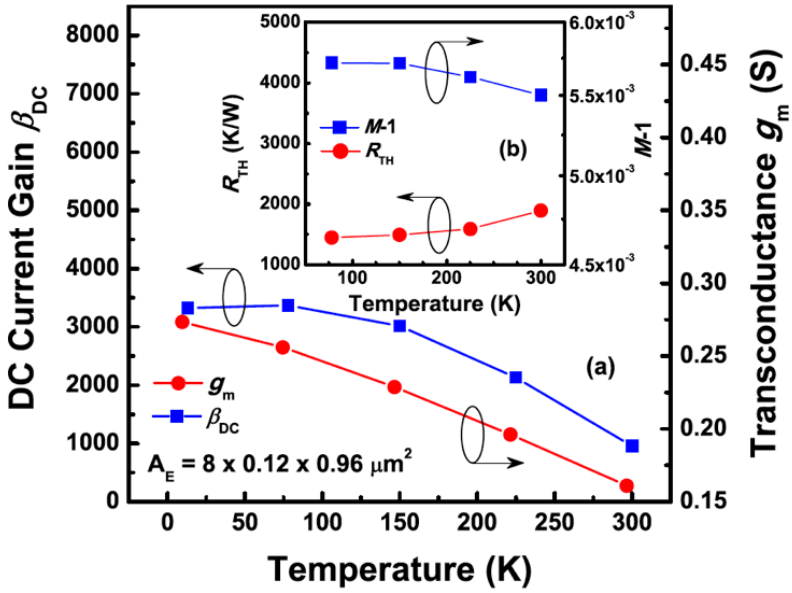

Figure 1.9: The DC current gain (β) and transconductance (g_m) of SiGe HBTs across temperatures. Inset: Thermal resistance and avalanche multiplication (M-1) of SiGe HBTs across temperatures (after [40]).

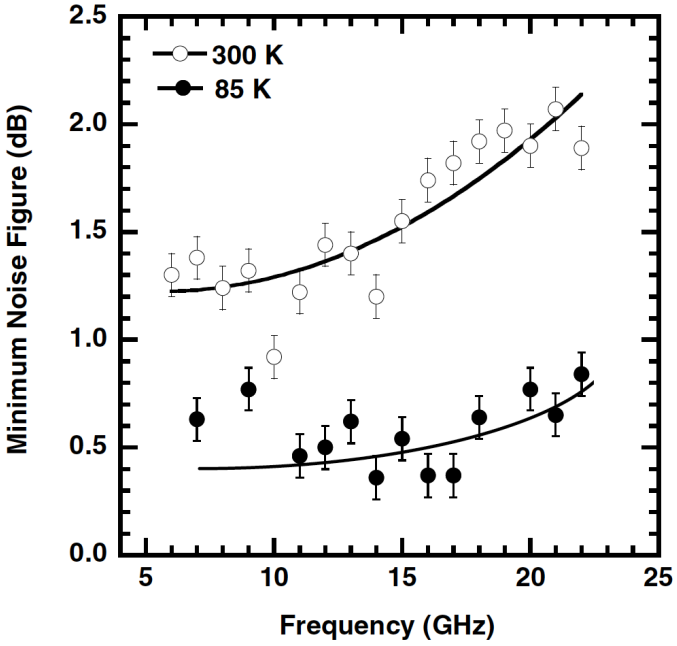

Figure 1.10: The noise figure of a SiGe HBT at 300 K and 85 K. (after [38])

putation are analyzed to adjust the control settings. Therefore, many bits of information for the control setting and the final results are processed or stored by the electronics. On the control side, high-resolution DACs require many digital transistors to convert digitally stored waveform to analog waveform with high fidelity while maintaining a wide range of tunability. Implementations of such digital blocks are naturally suited for CMOS devices due to the high density of CMOS devices per area. Experimentally, CMOS devices have demonstrated great performance in circuits targeting quantum computing applications [41, 20, 13].

In the readout stage, small RF signals need to be amplified and the phase of this signal relative to a reference RF signal is compared. These tasks require transistors with high speed and low noise. For both control and readout, RF mixers with low phase noise are needed. Both LNAs and mixers are routinely implemented using SiGe HBTs for telecommunication applications at room temperature, with impressive performance results. In addition, SiGe HBTs have proven performance when operated at deep cryogenic temperatures, with demonstrations at both device and circuit levels [42, 43, 44, 45, 46, 47, 48, 49]. Numerous SiGe LNAs have been implemented in SiGe HBTs with record performance in noise and power [50]. Overall, it is clear that both SiGe HBTs and CMOS have unique and complementary advantages for implementing electronics used in quantum science applications. Towards this goal, SiGe BiCMOS technology offers both types of devices on the same chip, enabling designers to pick the best from both worlds. This offers a unique, unrivaled performance that is irreplaceable by standard CMOS technology or III-IV technology. In summary, the ability to have dense CMOS circuits and high-performance SiGe HBTs on the same chip makes a compelling case for using SiGe BiCMOS technology for quantum science applications.

1.4 Research Objective

For CMOS devices, existing literature covers a wide range of topics including the electrical characteristics at cryogenic temperature, the operating physics, and the compact modeling of CMOS devices down to cryogenic temperatures [51, 52, 53]. However, there are few studies like this for SiGe HBTs, particularly at deep cryogenic temperatures of a few Kelvin. This is clearly an important missing topic if SiGe BiCMOS technology is to be used for quantum science applications. In particular, the following issues require further understanding or investigation, which are addressed in this thesis.

1. The lowest operating temperature of SiGe HBTs is unknown. The knowledge of this critically affects the system design, such as the optimal temperature stage to place SiGe HBT chips. It is unknown whether it is feasible to operate SiGe HBTs down to milliKelvin, which is common for quantum computing.

2. The transport mechanism of SiGe HBTs at cryogenic temperatures is unclear. This could lead to unexpected behaviors and reliability failures due to unforeseen physical mechanisms.

3. The existing compact model for SiGe HBTs are not suitable for a variety of design tasks demanded by quantum science.

4. For larger-scale circuits, the device to device variability could be an issue. This issue has been found to exist in CMOS devices operating at cryogenic temperature, but no literature exists for SiGe HBTs.

All of these issues are inter-related and require careful study. In this thesis, a systematic method is employed to investigate these issues. Figure 1.11 summarizes the four aspects of this research, along with the noted chapters that are tackling each issue. The deep cryogenic temperatures of milliKelvin to a few Kelvin is unknown territory for SiGe HBT devices.

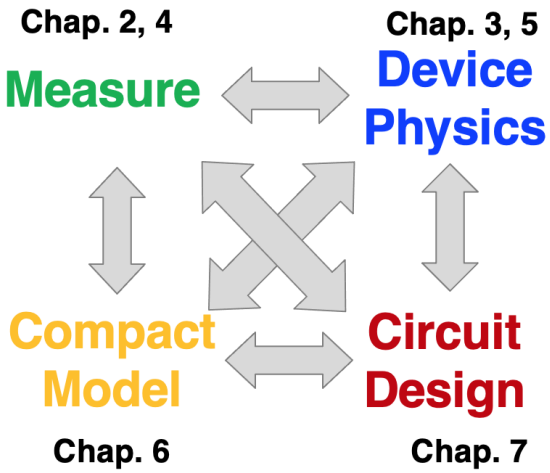

Figure 1.11: The four cornerstones of the research in this thesis, and the relationship between them. The chapters in this thesis are organized around the four concepts as noted near the corresponding labels.

The understanding of the operative device physics may reveal surprises not encountered in a typical design, such as a boost of performance, or a limitation that cannot be circumvented by traditional designs. In the end, comprehensive characterizations, good knowledge of the underlying physics, accurate compact models can all lead to optimized circuit design.

1.5 Thesis Organization

This thesis is organized as follows. Chapter 2 presents the characterization of SiGe HBT operation down to 70 mK. In Chapter 3, a method to differentiate between direct tunneling and quasi-ballistic transport is proposed, and the technology scaling trend is studied. The discovery of increased device-to-device variations of SiGe HBTs with lowering temperatures is presented in Chapter 4 while the physical origin of this increased variability is studied in Chapter 5. In Chapter 6, compact models of SiGe HBTs are presented along with the verification of the model using RF low-noise amplifiers. In Chapter 7, the effect of device variability on circuits is studied for current mirrors and RF amplifiers. Lastly, the summary and future work is presented in Chapter 8.

CHAPTER 2

CHARACTERIZATIONS OF SIGE HBTS DOWN TO 70 MILLIKELVIN

As explained in Chapter 1, it is important to know the operating boundary of SiGe HBTs at ultra-low temperatures. This chapter presents characterizations of SiGe HBTs down to a temperature of 70 mK, breaking the record of the lowest operating temperature of SiGe HBTs in literature. Transistor operation is shown at 70 mK, demonstrating the feasibility of using SiGe HBTs for quantum science application. In addition, the study shows that below about 40 K, the electrical characteristics of SiGe HBTs start to saturate. The root cause of this saturation phenomenon is investigated in detail in our earlier work in [54, 55], and summarized in Chapter 3.

2.1 Introduction

SiGe HBTs have long been recognized as a viable candidate for cryogenic temperature applications [42, 43, 44, 45]. When cooled, SiGe HBTs naturally exhibit improved frequency response (f_T and f_{max}), current gain, noise, bandwidth, and output conductance. Additionally, the compatibility with Si CMOS enables highly-integrated solutions to satisfy a wide variety of applications [39, 37, 36, 46, 49, 47, 35, 48].

An emerging field with interesting application opportunities for cryogenically-operated SiGe HBTs is quantum computing. Traditionally, qubit information is read out and sent through cables to room temperature electronics for amplification and processing. Due to the low signal-to-noise ratio (SNR), low bandwidth, and size inefficiency associated with this method, there are growing efforts to amplify readout signals at cryogenic temperatures with integrated electronics before sending it to room temperature [56]. Such electronics need to be operable at cryogenic temperatures and provide a large gain in order to boost the signal power and lower the SNR. In addition, it should have minimal noise (1/f and broadband)

and low power consumption (to avoid heating up and disturbing the qubits). Previously, CMOS devices, HEMTs, and MESFETs have been successfully demonstrated for such cryogenic preamplification tasks[57, 58, 59, 60, 61]. SiGe HBTs, however, are largely absent from this picture despite its known outstanding cryogenic capabilities. Recently, Sandia demonstrated a 10-100 times increase of SNR in qubit readout by using an off-the-shelf SiGe HBT for cryogenic preamplification [62]. The SiGe HBT in [62] is a discrete component from an early-generation SiGe technology with a room temperature f_{max} of 110 GHz. In comparison, current generation SiGe HBTs offer a much higher f_{max} under similar biasing conditions (as high as 500 GHz at 300K, and 800 GHz at 4.3K) [40, 63]. Besides a potentially higher speed for qubit readout and a higher achievable SNR due to higher gain, the newer devices can be operated at much lower bias currents (i.e., lower power) to achieve the same performance of the older devices, thanks to its overall higher f_T and f_{max}. Such flexibility for trading performance for power offers many advantages for cryogenic circuit design. Therefore, it is highly relevant to conduct the characterization of latest SiGe HBTs at mK temperatures.

This study presented the measurement data of a commercial fourth-generation SiGe BiCMOS technology [64]. Device measurements were performed from an ambient temperature of 300 K down to 70 mK. In addition, a magnetic field of $\pm$14 T was used to investigate non-ideal transport phenomenon. These are the first sub-100 mK cryogenic characterization results on a 4^{th} generation SiGe HBT, and the first time such a SiGe HBT is exposed to high magnetic field.

2.2 Device Technology and Measurement Setup

The SiGe HBT investigated in this work is from the GlobalFoundries' fourth-generation, 90-nm SiGe BiCMOS technology (GF 9HP), with a BV_{CEO} of 1.7 V and f_T/f_{max} of 300/360 GHz at 300 K. The devices measured have a drawn emitter geometry of 0.1x4.0 μm^2 [64].

Cryogenic measurements were made using a Quantum Design Physical Property Measurement System (PPMS) DynaCool® system with a dilution refrigerator (DR) insert, and an Agilent 4156C Semiconductor Parameter Analyzer. Test samples were die-attached on custom gold packages using indium solder to ensure good thermal conductivity. Electrical connections between the die and package were made by gold wirebonds of 1 mil (25.4 μm) diameter. The package is placed in the DR and all electrical connections inside the PPMS system are provided via low-loss superconducting aluminum twisted pairs. No cryogenic filtering was applied. Outside the system, the connections were adapted to triaxial cabling to minimize residual noise into the Agilent 4156C. During operation, the DR insert was placed in the bore of a 14 T superconducting cryomagnet and cooled to a base temperature of 50 mK.

The package temperature was constantly monitored to ensure minimal temperature fluctuations across all measurements. Due to the limited cooling power of the DR (0.25 μW at 100 mK), all measurements at 4 K and below were limited to a maximum collector current of 1 μA in order to maintain a stable DUT temperature. The only exception was to obtain a glimpse of the high-current operation at 70 mK and 100 mK, where an aggressive dc sweep was performed up to 0.1 mA collector current. This causes a transient temperature rise up to 50 mK above ambient (the ambient temperature begins rapidly changing at currents $>$ 1 μA). Higher current sweeps were avoided because the rapid heating can potentially damage the measurement system.

2.3 Results

In order to characterize the device across temperature, the Gummel characteristics were measured from 300 K to 70 mK, as shown in Figure 2.1. A zoomed-in view (i.e., V_{BE} = 0.85 V to 1.02 V) at 100 mK is provided in Figure 2.2(a). There are two observable non-idealities which are of interest. First, Figure 2.1 illustrates classical temperature scaling down to 16.7 K, but below 16.7 K, and all the way down to 70 mK, the curves (both I_C and

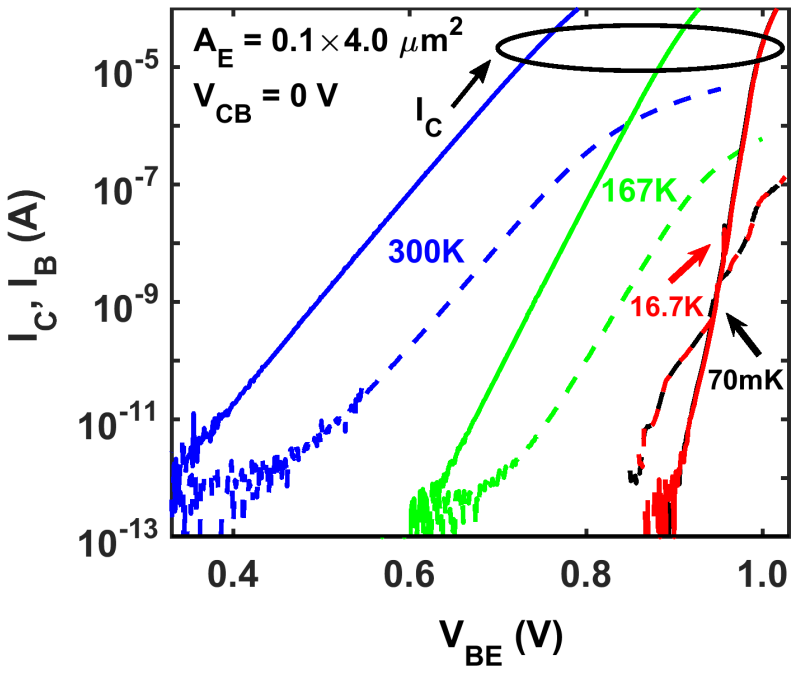

Figure 2.1: Measured Gummel characteristics of the 90-nm SiGe HBT at 300 K, 167 K, 16.7 K, and 70 mK. The Gummel characteristics at 16.7 K and 70 mK overlap and are virtually identical.

I_B) essentially overlap, even though the difference in reciprocal temperature is enormous (1000/T = 60 at 16.7 K vs. 1000/T = 14,286 at 70 mK). Second, the 16.7 K and 70 mK Gummel data from Figure 2.1 and the 100 mK Gummel in Figure 2.2(a) both show that the slope of the base current deviates from an ideal drift-diffusion 1/kT slope. In addition, two unusual sharp transitions or "steps" (denoted "lower step" and "upper step") in the base current can be observed. A total of 4 samples were measured but only one is shown in this study due to similarity.

To quantitatively examine the first non-ideality, the transconductance (g_m) and current gain (β) were plotted versus reciprocal temperature in Figure 2.3. Drift-diffusion theory dictates a linear 1/T dependence for $log(g_m)$, as shown in Figure 2.3 from 300 K to 40 K. Below 40 K, however, g_m flattens, indicating the slope of the collector is no longer steepening, possibly due to the quasi-ballistic transport through the neutral base and a higher effective electron temperature [65]. Similarly, the current gain flattens below 40 K. This

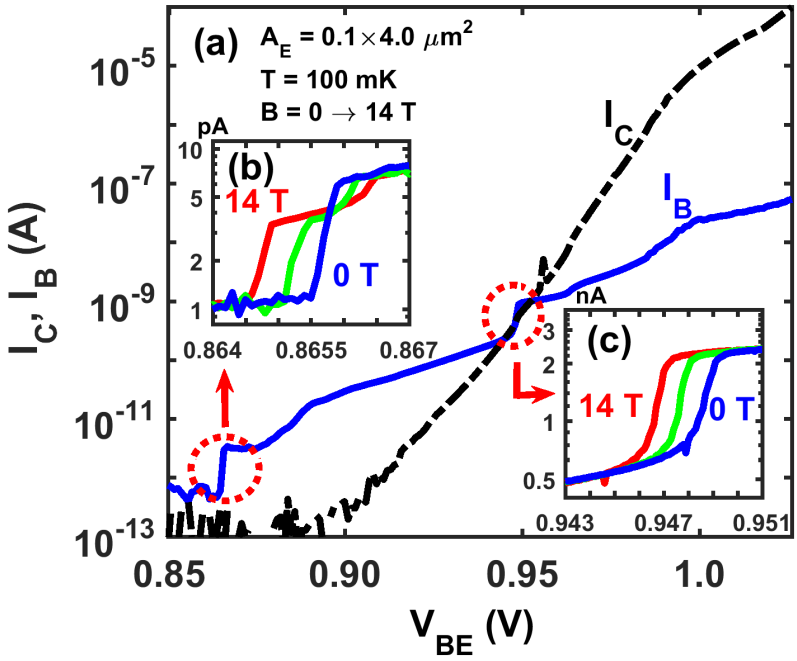

Figure 2.2: (a) Measured Gummel characteristics of the 90-nm SiGe HBT with V_{CB} = 0 V, at 100 mK. There are two "steps" in base current (solid line). (b) The lower step under three different magnetic fields at 100 mK showing a split. (c) The upper step in base current under three different magnetic fields showing a shift.

is consistent with the lack of temperature scaling between 16.7 K and 70 mK seen in Figure 2.1. As the reduction in temperature scaling is observed frequently, this characteristic can be considered a common feature of cryogenicallly-operated SiGe HBTs [65, 66, 67].

To investigate the base current non-ideality, a tunable static magnetic field (-14 T to +14 T) was applied the device at 100 mK. The reason for applying a magnetic field is two-fold. First, the temperature-independent Gummel characteristics imply that the transport mechanism should not have a strong temperature dependence. One possible mechanism which is consistent with this hypothesis is high-field tunneling in the emitter-base junction. Second, the sharp transition or step in current resembles the characteristic curve of electrons tunneling through a discrete energy level under an electric field [68]. According to the Zeeman effect, a degenerate energy level will split into two sub-levels under static magnetic field. Therefore, if the transport is initiated by tunneling, and the "steps" are caused by

discrete trap levels, an applied magnetic field should affect this transport mechanism and reveal itself in the magnetic field dependent Gummel characteristic.

Thus, a static and uniform magnetic field parallel to the transport direction (i.e., from emitter to collector) was applied to the SiGe HBT. The Gummel characteristics around the "step" region under different magnetic fields are shown in Figure 2.2(b) and Figure 2.2(c), for the lower and upper step, respectively. At 14 T, the lower step splits into two steps. As shown in the red curve in Figure 2.2(b), the lower portion of the current shifts to the left while the upper portion of the current shifts to the right, resulting in two discrete steps. This confirms that the tunneling is occurring through a discrete energy level, where the level splits into two non-degenerate levels under an applied magnetic field. At 7 T, the current lies between the 0 and 14 T curves, denoting that the current splitting has a linear dependence on applied magnetic field, which further confirms the presence of the Zeeman effect. Interestingly, the upper step does not show such a splitting, but rather a uniform shift in current under magnetic field. It is likely that when the discrete energy level splits, the upper sub-level is suppressed due to asymmetric transport rates [69]. Therefore, the discrete level essentially has a lower energy. This translates to current conduction at a lower base-emitter voltage, as shown in Figure 2.2(c). This step behavior has not been observed in cryogenic studies of earlier SiGe HBT platforms [46, 65]; however, as this step behavior is likely due to a high-field tunneling phenomenon in the emitter-base junction, the effect may be much weaker in less aggressively scaled technologies which naturally have lower doping levels. Alternatively, due to thermal broadening effects, the step behavior may be difficult to resolve and only reveal itself as the rich texture of base current at cryogenic temperatures [66]. Further investigation is needed to identify the dependency of this non-ideality on technology node and process conditions. In any case, all steps observed in this experiment (in all 4 samples) are below a few nA of collector current, and therefore, should have minimal effect on circuit applications.

Because we are ultimately interested in constructing readout circuits from these devices,

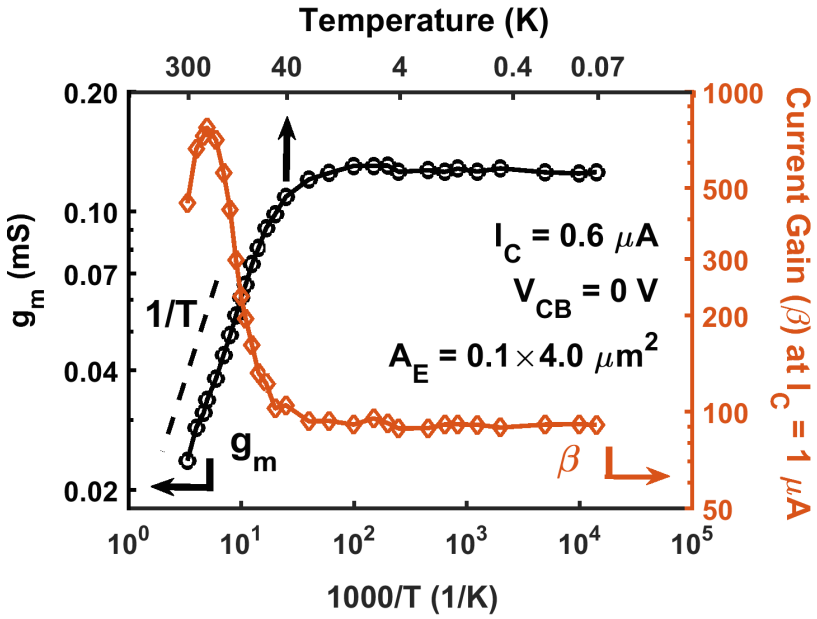

Figure 2.3: Plot of log transconductance (g_m) and current gain (β) across reciprocal temperatures. g_m is extracted at I_C = 0.6 μA and β is extracted at I_C = 1 μA.

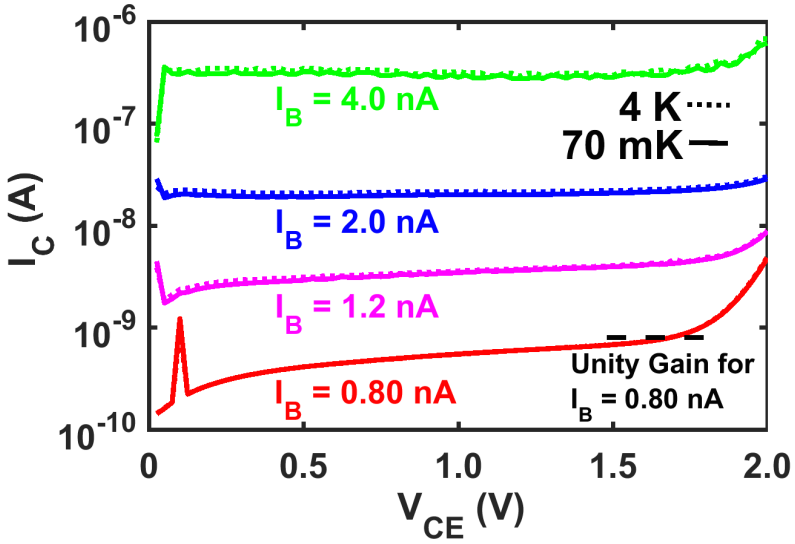

Figure 2.4: Output characteristics of the 90-nm SiGe HBT under different base current densities at 70 mK and 4 K.

we also measured the output family (I_C vs. V_{CE}). Since pre-amplification circuits are often implemented as transimpedance amplifiers in quantum computing applications, it is important to verify that the device exhibits high output resistance (i.e., a flat output characteristic) under a constant I_B driving mode. Additionally, the stability of device near the "step" region could be a potential concern for circuit design. To assess this, Figure 2.4 shows the fixed I_B output characteristics at 70 mK and 4 K under four different bias conditions. As shown, the output characteristics are indeed flat at very low base currents. The lowest curve also shows remarkable stability even though the I_B falls inside the upper "step" region. In addition, the output characteristics do not show much difference in collector current at the two temperatures. This, again, is welcome news from a circuit design perspective. In addition, the BV_{CEO} (not shown here) for I_E = 1 μA remained above 1.4 V from 4 K to 70 mK. From the temperature-invariant transconductance, it can be inferred that the device operation below 40 K should show minimal differences. This is clearly desirable from a cryogenic circuit perspective, since circuit designers can set an universal Q-point at 40 K and expect it to remain fixed down to 70 mK, provided self-heating is minimized.

2.4 Summary

We have characterized fourth-generation, 90-nm SiGe HBTs from 300 K to 70 mK. Interesting discoveries include temperature-invariant Gummel and non-ideal step effect in the base. Overall, the state-of-the-art SiGe HBTs provides promising operability for emerging deep cryogenic quantum computing applications.

CHAPTER 3

CRYOGENIC COLLECTOR TRANSPORT MECHANISMS OF SIGE HBTS

Given the saturation of electrical characteristics below a certain temperature, it is discovered from previous work in [54, 55] that direct tunneling is present in collector current transport in SiGe HBTs at cryogenic temperatures. This chapter first summarizes the transport mechanisms of collector current. Then, it proposes a method to distinguish between the quasi-ballistic transport and the direct tunneling based on the measurement of three generations of SiGe HBTs. Furthermore, this chapter investigates the effect of future technology scaling on direct tunneling.

3.1 Introduction

SiGe HBTs have recently gained attention due to their potential use in quantum computing readout circuits [62]. Quantum computing readout, together with other physics experiments such as charge sensing, detecting magnetic switching, or electron counting, creates a need for amplifiers that directly interface with quantum mechanical devices operating at deep cryogenic temperatures ($<$4 K, and even down to tens of mK) [70, 71, 72]. In order to connect the interior of the cryogenic refrigerators to room temperature instrumentation, experimental setups historically use lengthy cables that present significant loading on the already tiny output signals [56]. It was found that by adding a cryogenic pre-amplifier at mK temperatures, the system signal-to-noise ratio can be increased by as much as 100$\times$ [62]. Overall, these interface applications demand low noise, low to medium speed (up to several hundred MHz), and most importantly very low power. The cooling power of cryogenic refrigerators used for research purposes is typically on the order of a few Watts at 4.2 K and $< 1\ \mu W$ at 100 mK (as per vendor quotes from Lake Shore Cryotronics, Advanced Research Systems, Janis Research Company, Quantum Design, and [73]), while

that for commercial dilution refrigerators (DR) can be 0.5-1 mW at 100 mK [74, 75], increasing rapidly as the base temperature rises. In all cases, a significant portion of this power dissipation (the exact percentage is usually unknown) is used to cool mechanical fixtures such as sample plates, which leaves much less cooling capacity for the sample and interface circuits under study. Since the circuits are placed close to the samples (e.g., qubits or quantum mechanical structures) within the same cooling system, the limitations on cooling demands minimization of heat dissipation from the circuits in order to maintain the cold ambient temperature for the samples. For an optimistic scenario (100 mK using a commercial DR), interface circuits should at most operate in the range of 10-100 μW, if not less. For SiGe HBTs to successfully serve these emerging applications, it is important to understand the cryogenic transport mechanisms of SiGe HBTs at these extremely low power levels.

Previous studies have shown evidence for quasi-ballistic (non-equilibrium) transport and trap-assisted tunneling in early generations of SiGe HBTs [65, 46]. Recently, at cryogenic temperatures as low as 70 mK, SiGe HBTs exhibited non-classical characteristics in their measured currents, where a portion of the collector current was explained using a direct tunneling mechanism from emitter to collector [1, 54]. The same mechanism was discovered independently in a different SiGe technology, suggesting the mechanism is prevalent in advanced SiGe HBTs [76]. In light of this newly discovered mechanism, the present work proposes a simple method to differentiate direct tunneling mechanisms from quasi-ballistic transport. This provides new insight into how the technology scaling of SiGe HBTs affects the collector current in Gummel characteristics at cryogenic temperatures.

The present work also investigates the effects of generational scaling on the collector current of SiGe HBTs operating at cryogenic temperatures. Conventional scaling of SiGe HBTs leads to large improvements in peak unity-gain cutoff frequency (f_T) and maximum oscillation frequency (f_{MAX}), but such improvements are targeted for high-injection operation near room temperature [35, 40]. For low-power quantum interface applications at

cryogenic temperatures, it is crucial to know how temperature scaling affects the low-to-medium injection region due to the presence of non-classical transport mechanisms. Therefore, three generations of SiGe HBTs were measured and simulated in TCAD at cryogenic temperatures in order to examine the sensitivity of transport mechanisms to process parameters and provide insight for future profile designs.

3.2 Device Technology and Measurement Setup

The devices investigated in this work are from GlobalFoundries BiCMOS 5AM (0.5 μm), 8HP (130 nm), and 9HP (90 nm) technologies, and are referred to as 1^{st} [77], 3^{rd} [78, 79], and 4^{th} [80] generation devices. All devices presented are high-performance (HP) device variants from each generation; meaning, their collector profiles were optimized for maximum speed. The HP devices for the 1^{st}, the 3^{rd}, and the 4^{th} generations have peak f_T/f_{MAX} values of 50/65, 200/270, and 310/350 GHz, respectively, at 300 K. For simplicity, only NPN SiGe HBTs will be discussed. C-B-E-B-C layout structures, with an emitter length in the range of 1 - 10 μm were measured for each generation and the presented results were verified to be both reproducible and consistent across multiple samples.

The measurement was performed in a Quantum Design Physical Property Measurement System (PPMS) DynaCool® system. The system was configured to allow quick sweeps of temperature with a maximum cooling power of 5 mW at 1.5 K. Test samples were die-attached on custom gold packages using indium solder, and connected to the packages by gold wirebonds. The connections from the PPMS system were adapted to triaxial cabling to minimize residual noise and fed into the Agilent 4156C Semiconductor Parameter Analyzer for device characterization.

3.3 Existing Transport Theories

Three different transport mechanisms have been shown to exist in the collector current of SiGe HBTs operating at cryogenic temperatures: quasi-ballistic transport [65], trap-

assisted tunneling [46], and direct tunneling [54, 76]. A qualitative illustration of the three transport mechanisms with respect to bias is shown in the conduction band diagram in Figure 3.1. When the emitter-base voltage (V_{BE}) is small, no conduction is possible because carriers do not have sufficient energy to surmount the base potential barrier, and the base width is too large for any tunneling process. As V_{BE} is increased, however, the conduction band energy in the emitter is raised relative to the base. This effectively reduces both the base width and the barrier height for electrons in the emitter. If the base width is small and trap levels exist in the base, electrons can tunnel from the emitter into the trap states in the base and subsequently tunnel into the collector to generate transport current. With increasing V_{BE}, direct tunneling may become possible if the base width is small enough, such that electrons can tunnel directly (instead of via traps) from emitter to collector. At higher V_{BE}, electrons are brought close to the top of the base conduction band, and direct conduction occurs.

Direct conduction occurs via either drift-diffusion (in a thick base) or quasi-ballistic transport (in a thin base). For an extremely thin base, ballistic transport may occur [81, 82, 83]. Despite formal proof, quasi-ballistic transport was inferred to exist in SiGe HBTs and account for the low temperature non-ideal behavior in the collector current [65]. In quasi-ballistic transport, the collector current can be modeled phenomenologically as drift-diffusion, but with an electron temperature higher than the ambient temperature. At cryogenic temperatures, the lack of phonons results in less energy exchange between the electron ensemble and the lattice, which yields reduced cooling power for the electrons. This can explain the higher electron temperature. It is generally observed that the drift-diffusion 1/T scaling of collector current initially holds until the temperature is "low", at which point the current is better modeled with an effective electron temperature rather than an ambient temperature [66]. For this reason, in the present work, quasi-ballistic transport and drift-diffusion are assumed to be of the same form (both referred to as direct conduction), and the deviation of 1/T slope with cooling is assumed to come from quasi-ballistic transport.

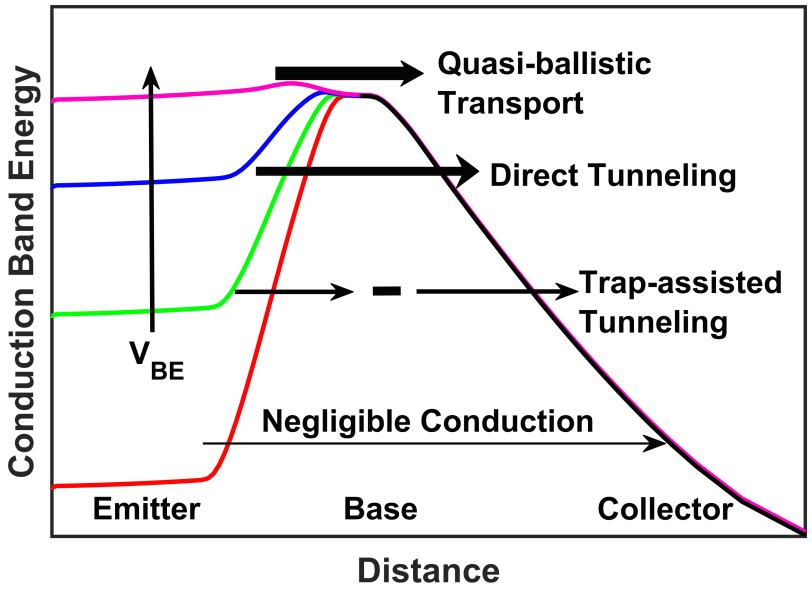

Figure 3.1: Qualitative illustration of various transport mechanisms from conduction band of SiGe HBTs under various V_{BE} values.

3.4 Direct Tunneling versus Quasi-ballistic Transport

The collector current density (J_C) versus V_{BE} of SiGe HBTs from three technology generations across temperature is shown in Figure 3.2. At high temperatures, the current is linear on a log scale (exponential) as drift-diffusion transport dictates. At both 1.8 K and 18 K, a non-ideal current at low injection is observed for all generations. In the 1^{st} generation device, the non-ideal current below 10 nA/μm^2 was shown to be driven by trap-assisted tunneling (TAT) [46]. Since a similar slope and location of the non-ideal current is observed in the 3^{rd} and the 4^{th} generation, it is plausible that the same TAT mechanism is present in the more advanced generations. In the present investigation, the region with a smaller slope, circled in Figure 3.2, is assumed to be due to a TAT mechanism. Section 3.5 will discuss the effects of scaling more in depth. For now, the main question to address is under which conditions the other mechanisms, namely direct tunneling and quasi-ballistic transport, exist, and more importantly, how to differentiate between them in the measurement.

To distinguish between the two, it is recognized that quasi-ballistic transport should

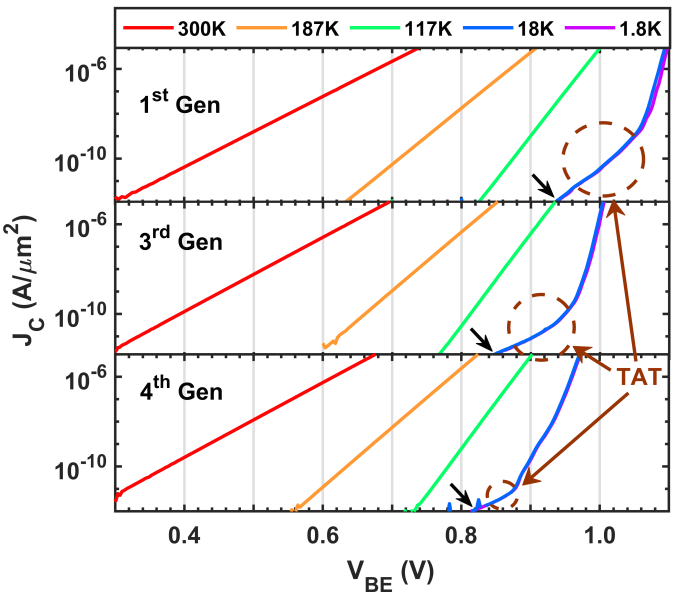

Figure 3.2: Measured collector current density versus V_{BE} of all three generations from 300 K to 1.8 K. Below 18 K, the curves overlap. The trap-assisted tunneling region (TAT) is circled, with its onset marked with arrows.

be only weakly dependent on the base width. In quasi-ballistic transport, the majority of carriers travel across the base without scattering. Therefore, a small change in base width does not change the already small scattering rate for carriers. In other words, the collector current component from quasi-ballistic transport should be invariant to changes in the base width. On the other hand, direct tunneling current is proportional to the transmission probability P for a rectangular base barrier, according to [84]

$$P \propto e^{-2W\sqrt{2m(U-E)/\hbar^2}} \tag{3.1}$$

where m is the effective mass of the electron, U is the potential energy, E is the energy of the electron ($E < U$ inside the potential barrier), and W is the barrier width. Although the base barrier is not rectangular, the exponential dependence still applies, and a change in barrier width (base width) is expected to result in an exponential change in tunneling current [54, 76]. Therefore, the difference between direct tunneling versus the quasi-ballistic transport can be revealed by measuring collector current for different base widths.

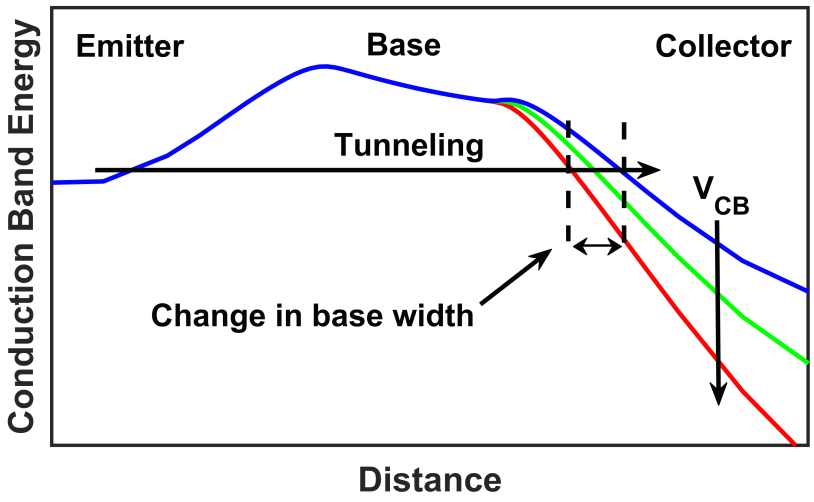

Figure 3.3: Qualitative illustration of conduction band diagram of SiGe HBTs showing the variation of base barrier width under multiple V_{CB} values.

To vary the base width without changing the other characteristics of devices, collector-base voltage V_{CB} is used as a tuning parameter. As shown in Figure 3.3, increasing V_{CB} shifts the collector conduction band energy (E_C) down, effectively reducing the tunneling barrier width. Similarly, decreasing V_{CB} increases the barrier width. Experimentally, the Gummel characteristics were measured at $V_{CB} = -0.3V$ and $V_{CB} = 0.5V$. As shown in Figure 3.4, apart from the increase of current at low injection due to band-to-band tunneling in the collector-base junction, the collector current of the 1^{st} generation device is invariant to changes in V_{CB}. However, a portion of the collector current in the 3^{rd} and the 4^{th} generation devices changes with V_{CB}.

To quantify this change in current, the normalized collector current at multiple V_{CB} values is plotted versus collector current density (J_C) in Figure 3.5. Clearly, in the 3^{rd} and the 4^{th} generation devices, collector current density from 10^{-11} to 10^{-7} $A/\mu m^2$ is very sensitive to V_{CB} and this sensitivity disappears towards higher J_C. In the 1^{st} generation, however, the sensitivity is much smaller throughout the J_C range. The strong sensitivity can be explained by the strong dependence of direct tunneling current on the base width (i.e., V_{CB}), while the weak sensitivity indicates quasi-ballistic transport dominates, which is independent of the base width. In other words, quasi-ballistic transport is present in all

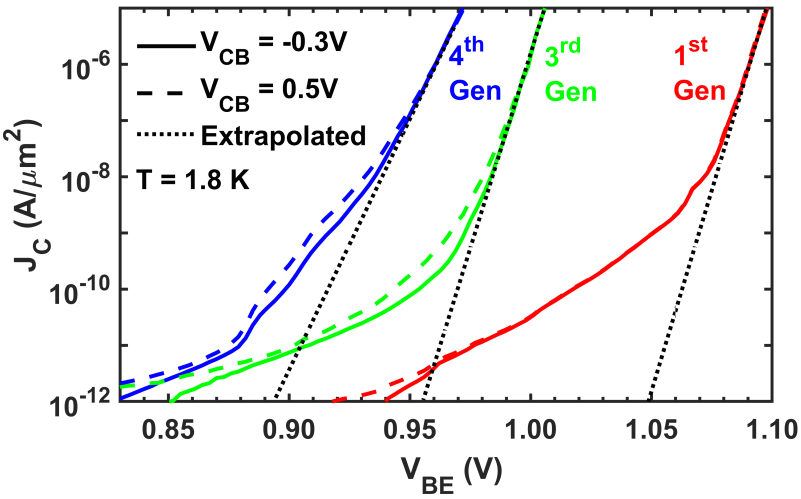

Figure 3.4: Collector current from the Gummel characteristics at 1.8 K for $V_{CB} = -0.3V$ and $0.5V$. Extrapolations of quasi-ballistic current (dotted lines) estimates the turn-on voltage for this mechanism.

generations, but direct tunneling is present only in the 3^{rd} and the 4^{th} generation devices. This makes intuitive sense.

There are two interesting observations to be made. First, it may seem surprising that the ratio of current is larger in the 3^{rd} than the 4^{th} generation device, although the 3^{rd} generation device should have less tunneling due to its larger base width. This can be understood, since the ratio of the current is proportional to the ratio of the tunneling probability, P. This ratio, from (Equation 3.1), is proportional to $e^{\Delta W}$, where ΔW is the change in base width. In other words, the ratio is proportional to the absolute change of base width, not its percentage change. Compared to the 4^{th} generation, the collector and base doping are lower in the 3^{rd} generation, causing a larger change in depletion width for the same change in V_{CB}. This causes the effective base width to change more significantly, resulting in a larger change in tunneling current for the 3^{rd} generation, as can be seen in Figure 3.5. The tunneling current, however, is still larger in the 4^{th} generation if we compare them at a fixed V_{BE} (barrier height), because of its smaller base width. The second observation is that direct tunneling can potentially degrade the device transconductance, g_m, as shown by the smaller slope of current in Figure 3.4. Therefore, the presence of direct tunneling requires

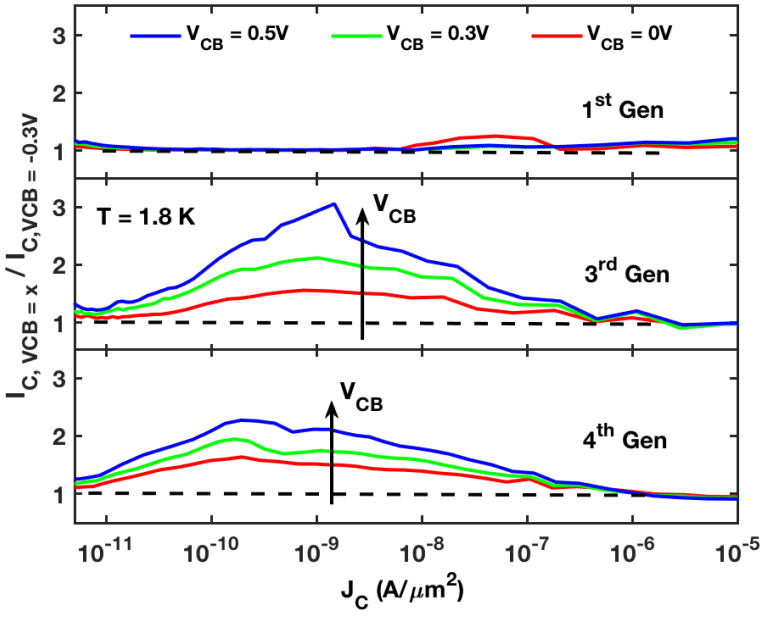

Figure 3.5: The ratio of collector current at various V_{CB} values to the current at $V_{CB} = -0.3V$ at 1.8 K in each generation.

more careful considerations for cryogenic circuit designs.

To gain more insight into how direct tunneling becomes significant as temperature decreases, the ratio of collector current between $V_{CB} = 0.5V$ and $-0.3V$ is plotted across temperature in Figure 3.6. At 82 K, the ratio is mostly constant for all generations, indicating that direct tunneling is negligible. As the temperature is lowered, a "hump" is progressively observable in the 3^{rd} and the 4^{th} generation devices, but not in the 1^{st} generation device. Since at a fixed V_{BE}, drift diffusion current decreases with temperature while the tunneling current remains roughly constant with temperature, tunneling becomes the dominant mechanism at low temperatures, thereby increasing the ratio. In particular, the hump due to direct tunneling first shows up in low injection and then slowly encroaches towards high injection as the temperature is lowered. When the drift diffusion current is replaced by quasi-ballistic current, the percentage of tunneling also stops changing, as shown in the 4^{th} generation device. For the 3^{rd} and the 4^{th} generation devices, the hump at 1.8 K corresponds to a $3\times$ and $2\times$ increase, respectively, which is very significant. In compari-

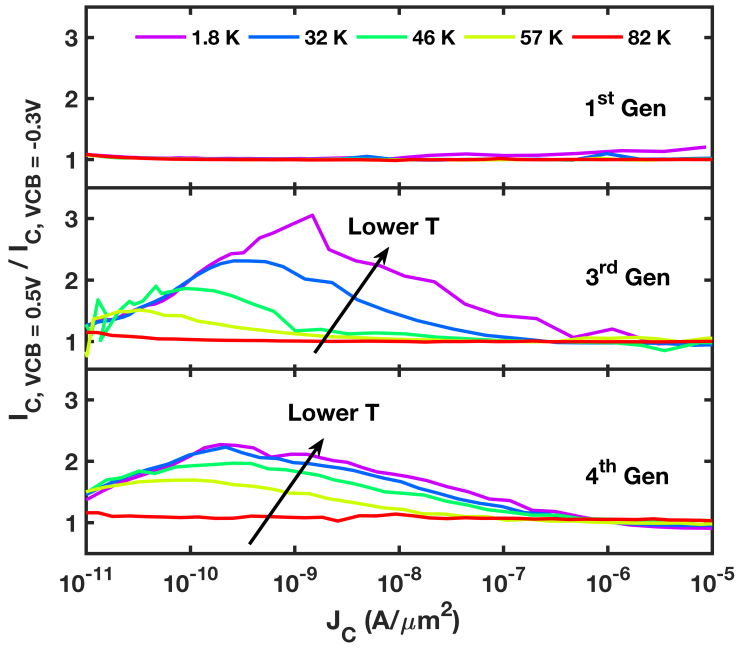

Figure 3.6: The ratio of collector current under $V_{CB} = 0.5V$ and $V_{CB} = -0.3V$ from 82 K to 1.8 K in each generation.

son, the ratio is constant in the 1^{st} generation, as the base width is too large for any direct tunneling. In short, direct tunneling can be observed in scaled technologies over a finite range of collector currents by simply varying V_{CB}.

3.5 Effects of Scaling on Transport

As SiGe HBT technology scales, there is a consequent decrease in base width, increase in collector doping, increase in Ge mole fraction, and increase in base doping. Aside from the increase in base doping, which raises the base potential barrier, all the other modifications will lower the base potential barrier. A lower barrier enhances the direct tunneling mechanism, which means if quasi-ballistic transport is not increased at the same rate, direct tunneling will become significant.

Quasi-ballistic transport is mostly dependent on the base barrier height. This can be understood by recognizing that most electrons do not have enough energy to overcome the

base barrier at cryogenic temperatures until the barrier height is close to zero, which is when the emitter quasi-Fermi level is similar in height to the base conduction band edge. The base barrier height is directly dependent on the integrated Ge content, which is indirectly related to the peak Ge content and the width of the Ge profile. As seen from the extrapolated line in Figure 3.4, the onset of the quasi-ballistic transport shifts towards smaller V_{BE} as technology scales, because the peak Ge content increases in the more advanced generations. To decrease the turn-on voltage of the quasi-ballistic transport, a larger Ge mole fraction is required. However, a larger Ge mole fraction requires a thinner base to maintain Ge film stability [35]. Additionally, technology scaling targets improved high frequency operation at room temperature, and typically does not exceed a 30% peak Ge content. Instead one must also shrink the base width using decreased thermal cycles to achieve the improved performance [85]. Therefore, the onset of quasi-ballistic transport is likely fixed if room temperature scaling rules are followed. The reduced base width, however, will increase the direct tunneling current relative to quasi-ballistic current. Such increase will be visible until quasi-ballistic current eventually rises above the direct tunneling current. In other words, direct tunneling is expected to dominate the collector current up to a higher V_{BE} in more scaled technology generations.

It is difficult to predict how the trap-assisted tunneling will change with technology scaling, since TAT, which relies on the presence of traps, depends on technology sensitive process steps, such as epitaxial growth conditions and tooling. For example, in [76], no TAT region is observed, which indicates that the specifics of the device structure design and processing conditions can play a major role in eliminating the TAT. In particular, a perimeter versus area (P/A) analysis can give more insight into the physical location of the traps in play. However, a P/A analysis is not meaningful in the present case because the standard device sizes in these highly scaled technologies allow for only a small range of P/A ratio (their emitter stripe width is fixed at minimum geometry and cannot be altered). That said, we can still estimate the energy of the trap levels based on the onset of the TAT from

the Gummel characteristics. As illustrated in Figure 3.1, quasi-ballistic transport occurs when the emitter quasi-Fermi level is close to the base conduction band edge. Since the onset of trap-assisted tunneling occurs earlier than the onset of the quasi-ballistic transport, the trap states must be located below the base conduction band. From Figure 3.2, the onset of trap-assisted tunneling is 0.87 V, 0.82 V, and 0.81 V for the 1^{st}, 3^{rd}, and 4^{th} generation devices, respectively. From Figure 3.4, the extrapolated onset of quasi-ballistic transport is about 1.05 V, 0.96 V, and 0.90 V. Therefore, TAT occurs about 100-200 mV before the onset of quasi-ballistic transport, indicating that the traps are likely located within 100-200 meV from the base conduction band edge. The good news here is that, as shown in Figure 3.4, in the 4^{th} generation devices, the TAT leakage current extends only to about 10 pA/um^2, much lower than in the 1^{st} generation devices, where it extends to 10 nA/um^2. This may be the result of both improved process technology and film tooling (yielding reduced trap density) and, equally importantly, an earlier onset of direct tunneling (as discussed above). From the data, the onset of direct tunneling current appears to shift to a smaller V_{BE} from the 1^{st} to the 4^{th} generations, thereby masking the trap-assisted tunneling. It is expected that, with scaling, such trends will continue and the direct tunneling will become more significant, causing the trap-assisted tunneling current to only appear at extremely low injection.

3.6 TCAD Simulations

We created three 2-D models in Sentaurus TCAD corresponding to three SiGe HBT generations and used them to investigate how scaling impacts the direct tunneling current. The models were fully calibrated to the room temperature DC and AC measurement data, with Slotboom bandgap narrowing and Phillips unified mobility models [86, 87]. To simulate tunneling current, the non-local tunneling model with Wentzel-Kramers-Brillouin (WKB) approximations was enabled, which does not alter the room temperature calibration. The non-locality of tunneling was incorporated as a generation/recombination rate dependent

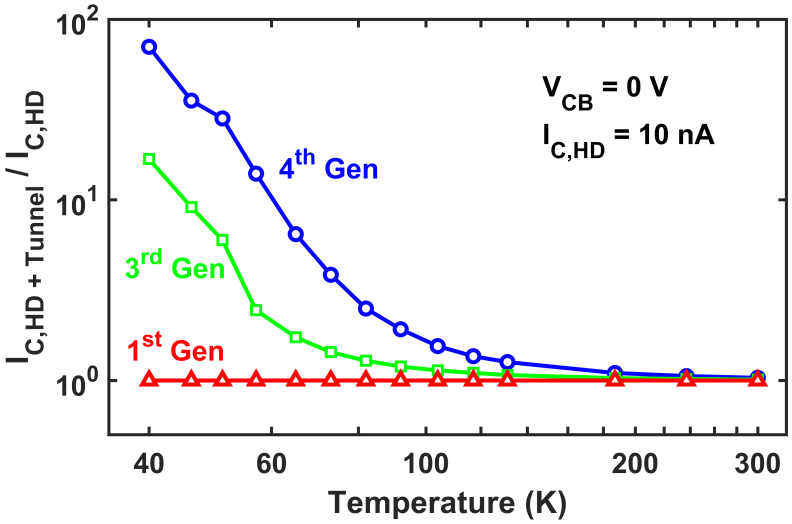

Figure 3.7: The ratio of collector current from hydrodynamic model with and without the tunneling model. The collector current is extracted from forward Gummel simulation at V_{BE} corresponding to I_C =10 nA in HD model

on the local quasi-Fermi level, electric field, and the potential profile along the tunneling paths [88]. The collector current was calibrated to the measurement results from 300 K to 40 K using only effective tunneling mass as a free parameter. Quasi-ballistic transport was not included in the TCAD due to the lack of physical models. For simulations below 50 K, recombination in the base was found to not affect the simulated collector current (as expected for direct tunneling).

Two investigations were conducted using TCAD. The first evaluated the amount of tunneling versus conventional drift-diffusion present in the transport current as the device scales. To evaluate this, the device was simulated with either hydrodynamic (HD) and tunneling models engaged, or with HD model alone. The ratio of collector current with and without tunneling at V_{BE}, corresponding to 10 nA of current from HD model alone, is plotted in Figure 3.7. As expected, the ratio of current with or without tunneling is constant throughout the temperature range in the 1^{st} generation device because the base width is too large for tunneling processes. For the 3^{rd} and the 4^{th} generation devices, however, the effects of direct tunneling on collector current begin to appear at as high as 180 K in the 4^{th} generation and 130 K in the 3^{rd} generation. Initially, the tunneling is simply a frac-

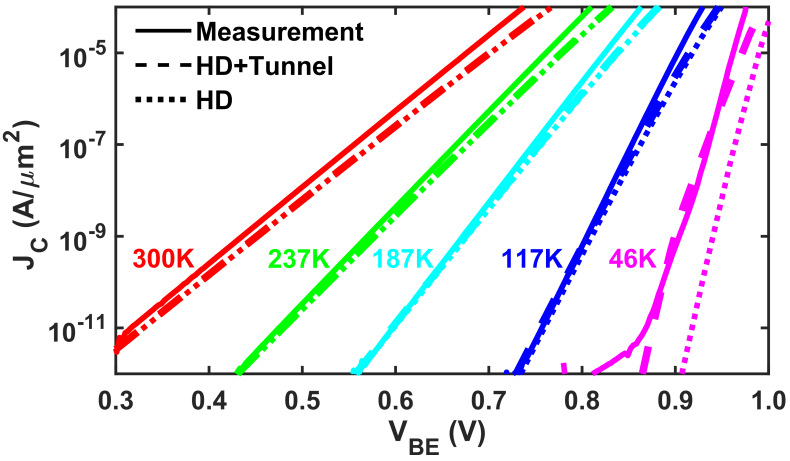

Figure 3.8: Collector current of the 4^{th} generation device from Gummel characteristics and TCAD simulations. TCAD simulations are in hydrodynamic model with or without the tunneling model.

tional increase compared to the hydrodynamic current. At lower temperatures, however, the hydrodynamic current becomes smaller compared to the direct tunneling current, and the ratio between the two increases dramatically. This can be visually seen in Figure 3.8, where the collector current from the simulation with and without the tunneling model, as well as the measured data, are plotted across temperature for the 4^{th} generation device. At high temperatures, the tunneling does not affect the collector current, and both simulations overlap with the data. At low temperatures, however, the simulated current with only the HD model can be as much as 2 to 4 orders of magnitude smaller than measured data! The HD model with direct tunneling, on the other hand, still models the data fairly well until a divergence above $1\ \mu A/um^2$ at low temperatures, as shown in Figure 3.8 for 46 K. The inaccuracies at high currents is due to the absence of a quasi-ballistic transport model, which is necessary to explain the continuous exponential increase in the measured current.

The second investigation evaluates the sensitivity of tunneling to the technology process parameters that can modify the base barrier shape. Four process parameters were chosen; namely, peak base doping, peak doping of the selectively implanted collector (SIC), peak Ge, and base width. In addition, a scenario where the base width is reduced while the base doping is increased by the same factor (to keep constant integrated base charge) was also

Table 3.1: Summary of Parameters in TCAD Simulation

Parameters	**Values**
Peak Base Doping	$7.8 \times 10^{19}\ cm^{-3}$
Peak SIC Doping	$9.0 \times 10^{18}\ cm^{-3}$
Peak Ge Fraction	0.27
Base Width	$16\ nm$

simulated. For comparison purposes, the base width is defined as the distance between the E-B and C-B metallurgical junctions. Though clearly multiple parameters will be scaled simultaneously in the real world (e.g., vertical reduction of base/collector profiles, reduction of emitter cap layer thickness, and modification of Ge profile, etc.), it is difficult to assess the contribution of individual parameters and make a fair comparison. Instead, the sensitivity to the various individual parameters was examined through the use of TCAD simulations. The base and SIC doping profile are assumed to be Gaussian shaped, and defined by the peak and standard deviation. The base width was adjusted by varying the standard deviation of the Gaussian boron profile, which changes the E-B and C-B doping intercepts. All simulations were performed using the calibrated 4^{th} generation TCAD model deck. The doping profile across the cutline in the intrinsic device is published in [54].

Shown in Figure 3.9 is the normalized collector current at a fixed V_{BE} versus the various process parameters, all simulated at 57 K. The process parameters were normalized to the values given in Table 3.1. Observe that scaling Ge affects the current density significantly, as even a slight reduction in Ge increases the base barrier width and height and thereby decreases the tunneling current exponentially. A separate simulation (not shown) reveals that at low temperatures the Ge content at the BE junction or the Ge grading in the base alone are much less important to the magnitude of tunneling than the total integrated Ge. To understand why, we can look at a typical Ge profile, which ramps up before the EB intercept and ramps down after the CB intercept. Compared to a graded Ge profile, a box Ge profile

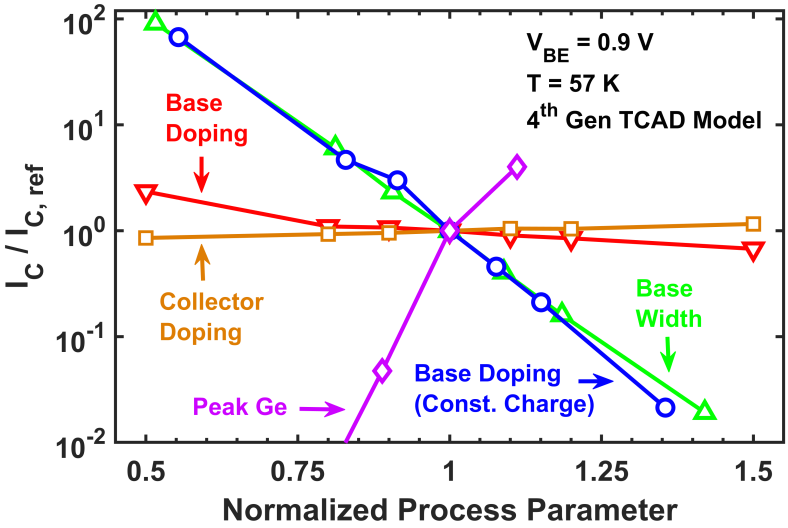

Figure 3.9: The simulated collector current versus process parameters that are individually varied in TCAD. Both the current and process parameters are normalized to the initial values.

reduces the base barrier height not only within the neutral base, but also in the EB and CB depletion regions, which effectively reduces the barrier width and exponentially increases the tunneling. Therefore, if a large collector current is desired (e.g., for a large current gain β), or if the turn-on voltage needs to be reduced, a larger Ge mole fraction (close to a box profile shape) throughout the base is preferred. There should be less concern over using a box profile in scaled technologies as the Ge film stability requirement would be more relaxed due to inherently smaller base widths. The base width reduction decreases the tunneling barrier width and allows significantly more collector current to flow. Even if the base doping is increased (higher barrier) while the base width is reduced, the current still increases significantly because the effect of base width dominates over the effect of base doping. The effect of collector doping is also small, which means that a higher collector doping to suppress Kirk and heterojunction barrier effects for the room temperature operation will not have much impact on cryogenic operation.

3.7 Summary

This study provides insight into cryogenic collector transport in SiGe HBTs. A unified picture of transport is summarized for three technology generations, where direct tunneling can be distinguished from quasi-ballistic transport through a simple experimental method. Among all transport mechanisms, trap-assisted tunneling could be potentially eliminated using optimized process technology, particular device structures, and improved film tooling with lower epi trap densities. At least, the industry scaling trend so far suggests that the effects of trap-assisted tunneling is diminishing in more advanced generations. Direct tunneling and quasi-ballistic transport are expected to become more dominant when the base width is scaled down, and the competition between the two transport mechanisms centers around the integrated Ge profile (width and Ge profile shape). Without an increase in Ge content, direct tunneling may dominate over quasi-ballistic transport to a higher collector current level. Through a process parameter sensitivity analysis using TCAD, total integrated Ge content and base width are determined to be the most important factors for optimizing cryogenic collector current in SiGe HBTs.

CHAPTER 4
VARIABILITY OF SIGE HBTS AT CRYOGENIC TEMPERATURES

During the study of Chapter 2, it was discovered that two nominally identical transistors from the wafer become more and more different as temperature decreases. This is concerning for circuit design and requires further investigation. This chapter studies the variability of SiGe HBTs down to a cryogenic temperature of 7 K. Different from similar variability characterizations for CMOS, this study characterizes not only the DC characteristics but also the RF characteristics. The results show that SiGe HBTs do have increased variability at lower temperature, but the DC variability is much larger than the RF variability. The root cause for DC variability is studied further in Chapter 5.

4.1 Introduction

Quantum computing has recently garnered increased attention in the cryogenic electronics community. In the current approach to realizing practical quantum computers, individual qubits are controlled through meter-long cables by equipment operating at room temperature. These cables occupy a large footprint inside the cryogenic chamber and demand large cooling power, which prevents further scaling of high qubit count quantum computers. A controller integrated circuit (IC) situated at cryogenic temperatures can alleviate this constraint by achieving the same functionality with a much smaller form factor [89]. Therefore, recent research efforts have focused on exploring suitable technologies for implementing monolithically integrated cryogenic controllers [41]. Among the various technologies, CMOS has been advocated as a candidate to leverage its mature manufacturing and scaling capability [90]. A technology not often considered in this context is SiGe BiCMOS technology, which can provide unique cost and performance trade-offs for cryogenic operation.

Among BiCMOS platforms, the availability of silicon-germanium heterojunction bipolar transistors (SiGe HBTs) can strongly complement CMOS in the analog/RF domain. SiGe HBTs are known to have improved bandwidth and lower noise at lower temperatures [39, 35]. Additionally, similar performance vs. CMOS devices can usually be obtained from SiGe HBTs at a less advanced lithographical node, thereby reducing the cost. Therefore, the implementation of cryogenic controllers using SiGe BiCMOS is worth exploring. As a first step towards this goal, previous work has verified the DC operation of SiGe HBTs down to temperatures as low as 70 mK [1].

Apart from the DC operability at cryogenic temperatures, however, the process variation of any cryogenic technology needs to be characterized to enable practical circuit designs. Without knowledge of device variability, fabricated circuits can fall outside the acceptable range of performance, which can reduce the overall yield, and increase production costs [91]. In particular, variations in DC characteristics affect the biasing of circuits, while variations in RF characteristics affect the bandwidth and noise, all of which are important for cryogenic controller design. In CMOS devices, increased variability of DC characteristics has been observed at cryogenic temperatures [92]. DC variability of SiGe HBTs at low injection levels has been studied in our previous work, but thus far, no studies have investigated the variability of SiGe HBTs at high injection or at high frequency, which are valuable for any RF amplifier designs.

The goal of the present work is to explore the variability of DC and RF parameters across temperature, down to as low as 7 K (the lower limit of our RF test system). The data reveal an increased variability for most DC and RF parameters as temperature decreases. This is the first work that demonstrates such increased variability of SiGe HBTs below $<$10 K. In addition, the implications for such increased variability on cryogenic circuits, as well as possible mitigation methods are discussed.

4.2 Device Technology and Measurement Setup

The SiGe HBTs investigated in the present work were fabricated using the GlobalFoundries 90-nm SiGe BiCMOS technology (GF BiCMOS9HP). These devices feature a BV_{CEO} of 1.7 V and f_T/f_{max} of 310/350 GHz at 300 K [64]. The samples were fabricated from the same batch of wafers on a multi-project wafer run. DC and high frequency measurements were performed on 14 samples at 78 K and 300 K, and five samples at 7 K. All devices were configured as high-performance CBEBC devices with a emitter geometry of 0.1×4.0 $\mu\mathrm{m}^2$.

Measurements were made in Desert Cryotronics® TTP6 open-cycle liquid helium system using DC to 50 GHz GSG probes. At each temperature, short-open-load-thru (SOLT) calibration was performed, followed by on-die short and open de-embedding. AC measurements from 0.5 to 50 GHz were performed using an Agilent E8363B Vector Network Analyzer, and DC measurements were obtained using an Agilent 4155C Semiconductor Parameter Analyzer.

4.3 Results

The Gummel characteristics (collector current, I_C, and base current, I_B, versus base-emitter voltage, V_{BE}) are shown in Figure 4.1 for 300 K, 78 K, and 7 K, at V_{CB} = 0 V. The error bars denote the minimum and maximum measured current at each bias voltage, while the lines show average values. The error bars indicate that the variability increases with lower temperatures. To quantify the amount of variability, relative standard deviation (RSD) is introduced as a figure-of-merit and is used throughout the present work. RSD is defined as the standard deviation (σ) divided by the mean (μ) of a parameter. The RSD of I_B and I_C in the Gummel characteristics is plotted in the inset of Figure 4.1, for all temperatures. From the plot, two observations can be made:

First, both I_C and I_B show increasing variability from 300 K down to 7 K. For example,

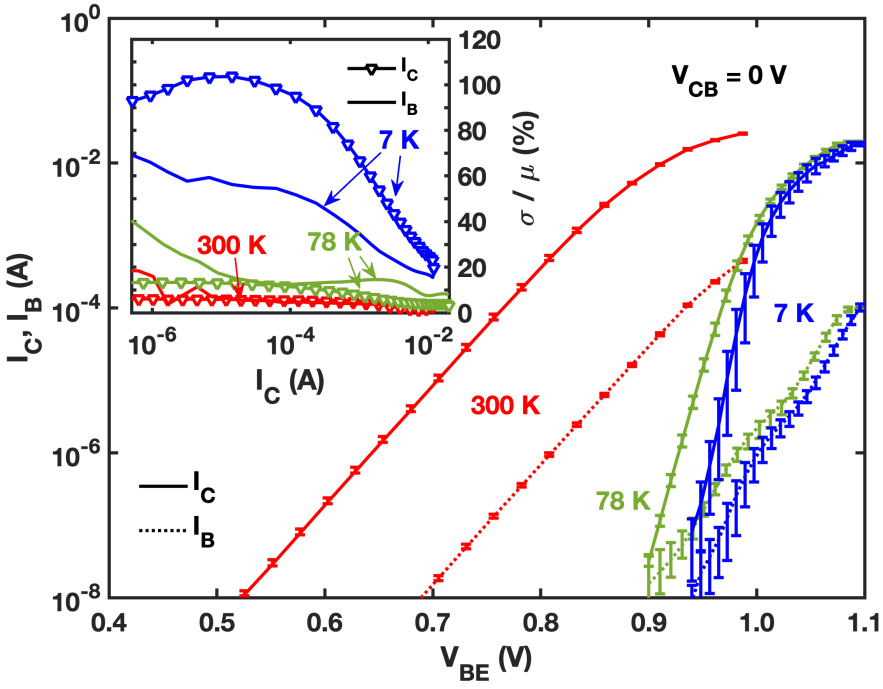

Figure 4.1: Measured Gummel characteristics at 300 K, 78 K, and 7 K for V_{CB} = 0 V. The error bars denote the minimum and maximum measured current at each bias voltage. Inset: RSD of I_B and I_C as a function of average measured I_C at each temperature.

when the average measured I_C is 1 mA, the RSD of I_C increases from 5% at 300 K to 67% at 7 K. Possible reasons for this large increase are variations in doping, the Ge profile, and the electric field, as discussed in our previous work [4]. Second, I_C has decreased variability in high injection level ($>$1 mA), and this trend is increasingly noticeable at lower temperatures. High injection Gummel characteristics are affected by series resistance, which is expected to have less variability than the transport currents (exponentially dependent on temperature and voltage). Self-heating may contribute to the reduced variability at high injection as well, due to the trend of reducing variability at higher temperatures. However, neither effect is present at low injection and does not contribute to the increased variability there. The fact that the high injection current sees less variability has important implications for the choice of circuit bias, which is discussed below.

Two additional important DC parameters are the DC current gain (β) and the transconductance (g_m, the derivative of I_C vs. V_{BE} in the Gummel characteristics). As shown in

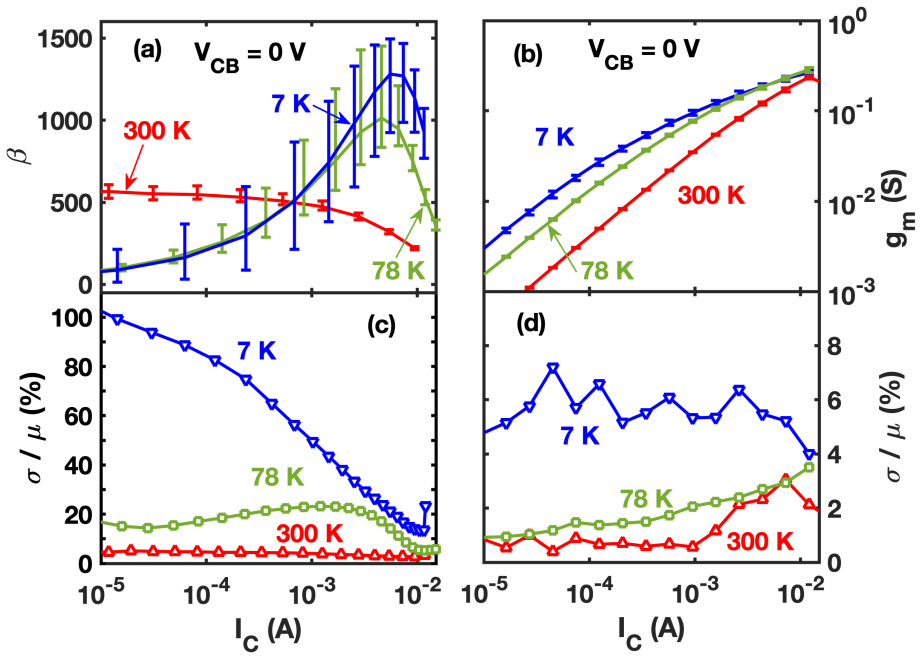

Figure 4.2: (a): Measured β versus I_C at 300 K, 78 K, and 7 K for V_{CB} = 0 V. The error bars denote minimum and maximum beta at each bias. (b): Measured g_m versus I_C at each temperature for V_{CB} = 0 V. (c): RSD of β as a function of I_C. (d): RSD of g_m as a function of I_C.

Figure 4.2(a), β is larger at lower temperatures, as expected, with a peak β increasing from 565 at 300 K to 1288 at 7 K. From 300 K to 7 K, g_m for fixed I_C increases by 7x at low injection and 1.2x at high injection, as shown in Figure 4.2(b). While larger β and g_m are often desirable for circuit design, their variability also increases at lower temperatures, as shown in Figure 4.2(c) for β and Figure 4.2(d) for g_m. The RSD of β increases from $<$5% at 300 K to around 20% at 78 K and close to 100% at 7 K. The increased RSD of β mainly comes from the increased RSD of I_C and I_B. On the other hand, the RSD of g_m is much smaller than that of β for most range of I_C.

Typically, the gain of amplifiers are affected by either β or g_m, depending on the topology. β variation is more relevant for current amplifiers, while g_m is more important for voltage amplifiers. One design challenge is how to ensure the stability of the amplifier in the presence of large gain variation. Since qubit signals are small, cryogenic controllers would benefit from high gain amplifiers, though they are then more prone to stability issues.

If the amplifier is designed to be stable with the highest gain, post-fabrication chips with lower gain may not meet design specifications. If designed around the mean value of gain, post-fabrication chips with higher gain may become unstable. Therefore, the variability of β or g_m needs to be carefully evaluated for each design.

Apart from DC parameters, RF characteristics strongly affect circuits as well, but their variability at cryogenic temperatures has not been investigated previously. The unity current gain frequency (f_T) was extracted across I_C and plotted in Figure 4.3(a). Representative fitting curves are plotted in Figure 4.3(b), where the extraction fits the deembedded h_{21} versus frequency with a line of -20 dB/decade slope. Together with the increase of average peak f_T (from 315 GHz at 300 K to 549 GHz at 7 K), the RSD also increases, as shown in Figure 4.3(c). At low injection, the RSD increases from 5% at 300 K to around 25% at 7 K. At high injection, the variability is similar between all three temperatures. The RSD at 300 K is relatively flat across I_C, while higher variability is observed at lower injection at 7 K. To obtain added insight into the variability of f_T, we can express f_T to the first order as

$$\frac{1}{2\pi f_T} = \tau_{EC} + \frac{kT}{qI_C}C_t \tag{4.1}$$

where τ_{EC} is the total emitter-collector transit time and C_t is the sum of base-emitter capacitance (C_{BE}) and base-collector capacitance (C_{BC}). Therefore, it is useful to decompose f_T and extract τ_{EC} and C_t.

τ_{EC} was extracted from $1/(2\pi f_T)$ versus $1/I_C$ at each temperature and is shown in Figure 4.4(a), with typical extraction curves in Figure 4.4(b). The average τ_{EC} decreases at lower temperatures because of less carrier scattering and the higher mobility of carriers. At V_{CB} = 0 V, τ_{EC} decreases from 0.42 ps at 300 K to 0.26 ps at 7 K. As plotted in Figure 4.4(c), the RSD of τ_{EC} increases from 2.8% at 300 K to around 7.2% at 7 K.

C_t was extracted from the linear slope of the fitting curves in Figure 4.4(b) and plotted in Figure 4.4(d). The RSD (not shown) of C_t at V_{CB} = 0 V increases from 2.2% at 300 K, to 9.0% at 78 K and 10.7% at 7 K. Since C_t consists of C_{BC} and C_{BE}, we can extract these

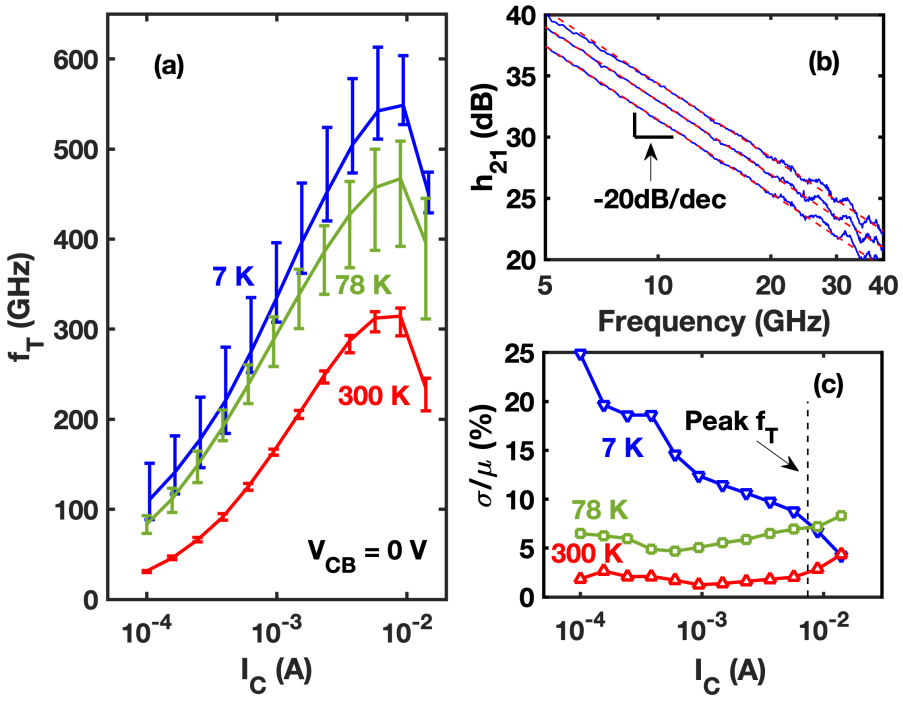

Figure 4.3: (a): Extracted f_T versus I_C at 300 K, 78 K, and 7 K. (b): Representative extraction curves of f_T from h_{21} for different V_{BE} bias. (c): The RSD of f_T versus I_C at each temperature.

two parameters for additional information.

One portion of the total C_t comes from C_{BC}, which is extracted at $V_{BE} = 0$ V from the imaginary part of $-(y_{12})/\omega$ (ω is the angular frequency). As shown in Figure 4.5(a), the average C_{BC} at V_{CB} = 0 V decreases due to the increased junction built-in voltage and depletion width with cooling. While the value of C_{BC} decreases, the RSD of C_{BC} increases from 0.5% at 300 K to 3.6% at 7 K, as shown in Figure 4.5(b).

The other portion of C_t is C_{BE}, which can be further divided into depletion capacitance (at low V_{BE}) and diffusion capacitance (at high V_{BE}). The depletion C_{BE} is extracted from the imaginary part of $(y_{11} + y_{12})/\omega$ at V_{BE} = 0 V for 300 K and 78 K, and plotted in Figure 4.5(c). Although the 7 K data are not yet available, the data at 300 K and 78 K shows that depletion C_{BE} decreases from 300 K to 78 K, while the variability increases from ≈4% at 300 K to ≈20% at 78 K, as shown in Figure 4.5(d). Overall, C_{BE} and C_{BC} has increased variability at lower temperatures, which causes increased variability of C_t

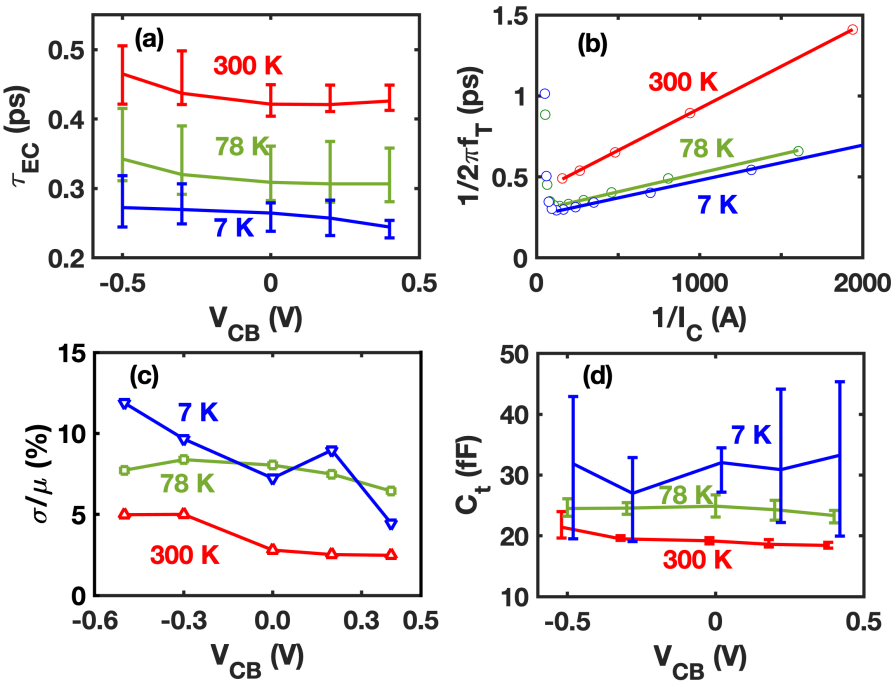

Figure 4.4: (a): Extracted τ_{EC} versus V_{CB} at 300 K, 78 K, and 7 K. (b): Typical extraction curves of τ_{EC} from $1/(2\pi f_T)$ versus $1/I_C$. (c): The RSD of τ_{EC} versus V_{CB} at each temperature. (d): Extracted C_t versus V_{CB} at each temperature.

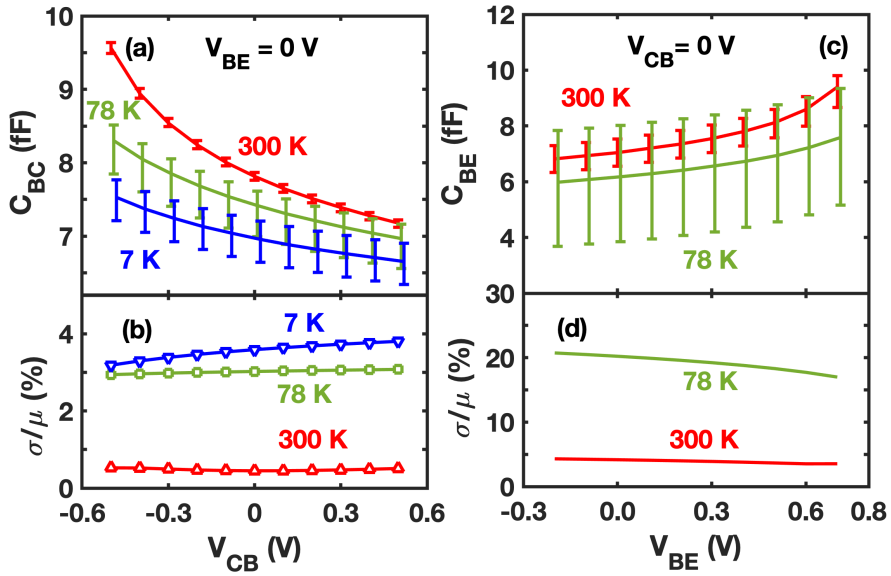

Figure 4.5: (a): Extracted C_{BC} at V_{BE} = 0 V at 300 K, 78 K, and 7 K. (b): The RSD of C_{BC} versus V_{CB} at each temperature. (c): Extracted depletion C_{BE} versus V_{BE} at V_{CB} = 0 V at 300 K and 78 K. (d): The RSD of depletion C_{BE} versus V_{BE} at each temperature.

and f_T.

Table 4.1: Summary of Measurement Results

Parameters	**300 K**	**78 K**	**7 K**
I_C(mA)	$1 \pm 5\%$	$1 \pm 48\%$	$1 \pm 55\%$
$I_B(\mu A)$	$10 \pm 5\%$	$10 \pm 38\%$	$10 \pm 17\%$
Peak β	$566 \pm 4.6\%$	$1013 \pm 17\%$	$1281 \pm 21\%$
β@I_C=0.7 mA	$500 \pm 4\%$	$515 \pm 23\%$	$507 \pm 56\%$
g_m(mS)@I_C=1 mA	$35 \pm 0.6\%$	$77 \pm 2.1\%$	$95 \pm 5.3\%$
Peak f_T(GHz)	$315 \pm 2.6\%$	$467 \pm 7\%$	$549 \pm 7.1\%$
f_T(GHz)@I_C=0.1 mA	$31 \pm 1.8\%$	$83 \pm 6.5\%$	$111 \pm 24.9\%$
τ_{EC}(ps)@V_{CB}=0 V	$0.42 \pm 2.8\%$	$0.31 \pm 8.0\%$	$0.26 \pm 7.2\%$
C_t(fF)@V_{CB}=0 V	$19.4 \pm 2.2\%$	$22.8 \pm 9.0\%$	$32.0 \pm 10.7\%$
C_{BE}(fF)@V_{BE}=0 V	$7.0 \pm 4\%$	$6.2 \pm 20\%$	/
C_{BC}(fF)@V_{CB}=0 V	$7.7 \pm 0.4\%$	$7.4 \pm 3.0\%$	$7.0 \pm 3.6\%$

A summary of the measurement results and their RSD is given in Table 4.1. Overall, all DC and RF parameters show increasing variability at lower temperatures. In particular, DC parameters tend to have more variability than RF parameters (except g_m). This increased variability presents new challenges for cryogenic circuit designers.

4.4 Discussion

The classical trade-off between power consumption and bandwidth will inevitably apply to cryogenic controller design. In particular, cryogenic controllers need to dissipate as little power as possible to prevent heating of nearby qubits and shortening their coherence time [62]. The scaling of quantum computers to contain more qubits also requires more controller channels, demanding less power per channel to accommodate the same (constrained) cooling power from the cryostat [56]. On the other hand, higher bandwidth can enable faster readout with higher fidelity, but requires transistors to be biased at high injection and dissipate additional power. This creates a trade-off between lower power and

higher bandwidth.

The present work demonstrates an undesirable additional trade-off at cryogenic temperatures. With less power, transistors have more variability in DC and RF characteristics. Based on the variability observed in the Gummel characteristics in Figure 4.1, biasing transistors at lower injection levels creates more mismatch in current between different samples. Coupled with the variability in β and g_m for the same current, this mismatch causes the gain to be different between transistors. Such mismatch could affect important circuit building blocks such as current mirrors and differential pairs. Apart from the concern for potentially degraded amplifier stability, unequal gain between amplifiers could affect large-scale system design. In cryogenic controllers, qubit-generated signals must be amplified and compared to a reference level for digitization. An unequal gain in either amplifier gain or reference level can create erroneous outputs. Therefore, this requires special attention to mitigate the increased variability.

To potentially mitigate the variability we have presented, one can bias the transistors at higher current levels, or have more transistors in parallel to average the mismatch. The optimal number of parallel transistors will likely prove to be a design parameter for each cryogenic application. Circuit feedback can also be used to reduce the circuit sensitivity to device mismatch. It will likely prove important to have well-defined corner models that enable designers to assess these circuit variations before fabrication. Overall, this increased variability seems to be common among both CMOS and BiCMOS technologies (i.e., they are fundamental), but successful demonstrations of cryogenic controllers in CMOS prove that overcoming such problems in BiCMOS technology is possible. Despite possible mitigations, it is imperative that designers are aware of the increased variability during the design process [41].

4.5 Summary

We have characterized commercially-available 90-nm SiGe HBTs at 300 K, 78 K, and 7 K. Overall, the variability of DC parameters increases significantly with cooling into the deep cryogenic regime. In particular, DC parameters, including bias and gain, exhibit larger variability at lower temperatures. If left unchecked, this increased variability can shift circuit performance outside specifications. The RF parameters also have increased variability at lower temperatures, though the variability is generally less than that of DC parameters. At each temperature, operation at higher injection levels tends to mitigate variability, at the expense of power. On a positive note, DC and RF performance is strongly enhanced with cooling in SiGe HBTs. As long as circuit designers are aware of the emerging challenge of increased variability with cooling, and carefully account for it, it should be possible to overcome this challenge.

CHAPTER 5
PHYSICAL ORIGIN OF VARIABILITY IN PN JUNCTIONS AND SIGE HBTS

Given the variability results in Chapter 4, it is of interest to investigate why SiGe HBTs have increasing variability at lower temperatures. This chapter uses TCAD to examine the relevant operating physics of SiGe HBTs and arrives at a unified picture of variability across temperature including not only SiGe HBTs but also more fundamentally, the PN junctions. The presence of mathematical terms in the exponent of the Arrhenius function is one of the culprits for increasing variability at lower temperatures. Experimental results show that mechanical stress from the manufacturing process plays an important role in creating and enhancing the variability at cryogenic temperatures.

5.1 Introduction

SiGe HBTs have long been a strong candidate for cryogenic electronics [39]. Recently, increasing effort has been spent on leveraging the improved operating speed and lower transistor noise of SiGe HBTs at cryogenic temperatures for emerging applications such as quantum computing and quantum communications. Other relevant cryogenic applications include radio astronomy and deep space exploration [35]. In all cases, a solid understanding of the SiGe HBT and its operative mechanisms at cryogenic temperatures is desirable.

Our preliminary work in [3] shows that, for the same set of samples, the device-to-device variation of the terminal current in SiGe HBTs at a fixed bias voltage increases with decreasing temperatures. Such an increase leads to larger uncertainties in bias point, amplifier gain, data converter linearity, resulting in system-level degradations [93]. To mitigate these effects, more elaborate built-in tests and calibration routines are required, which implies a larger chip area and more power dissipation, both undesirable [94]. To date, no literature comprehensively explains why the variability of SiGe HBT parameters

increases at cryogenic temperatures, which is the motivation for the present work. It is worth noting, however, that other technology candidates, such as CMOS, also exhibit larger parameter variability at low temperatures [92]. Therefore, SiGe HBTs still have advantages over other technology candidates. With sufficient understanding, it should be possible to devise variability mitigation strategies and realize high-performance cryogenic circuits and systems using SiGe HBTs.

In the present work, we analyze several physical mechanisms that can potentially lead to variability in SiGe HBTs, supplemented with quantitative results using calibrated TCAD simulations. Further, it is shown that variability can exist in not only SiGe HBTs but also other minority-carrier devices such as PN junctions. It is expected that this universal picture of parameter variability will be helpful for the future development of SiGe HBTs and other minority-carrier devices intended for cryogenic operation.

This rest of this chapter is divided into five sections. Section 5.2 explains the operative mechanisms responsible for variability across temperature. Section 5.3 presents TCAD simulations to illustrate the effects of various physical mechanisms on PN junction variability, as well as the effect of Ge profile and tunneling mechanisms on SiGe HBT variability. Section 5.4 compares numerical simulations with measurements to provide additional insight into the sources of variability. Section 5.5 discusses possible mitigation methods. Section 5.6 concludes this work.

5.2 Analysis of Variability

The sources of variability in SiGe HBTs can be investigated at different levels, as shown in Figure 5.1(a). At the process level, variability can be traced to factors such as differences in processing temperatures (deposition, annealing), implantation energies, process uniformity across wafer, etc., all of which translate to device-level variability in doping, stress, and Ge profiles, as well as device dimensions [95]. These differences result in the physical change of key material properties that can affect transport mechanisms and therefore the

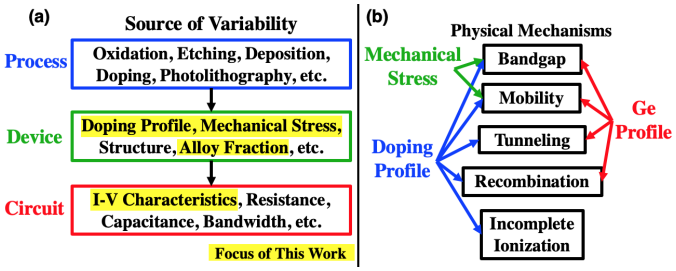

Figure 5.1: (a) Summary of variability sources at different levels. (b) Mechanisms and sources of variability discussed in the present work.

terminal current-voltage (I-V) characteristics of the device. To simplify the discussion, the three goals of the present work are summarized here. First, only the effects of doping, stress, and Ge profile are investigated. The variability of profiles always refers to the final (aggregated) variability, after all process steps are completed. Second, we are interested in how the profile variability affects the terminal current, in particular, at current levels far below the onset of high injection effects and self-heating. Third, different from that of typical variability studies (e.g., in [96]), the focus of the present work is on the temperature trends of variability, from room temperature down to deep cryogenic temperatures.

Variability analysis often requires numerical methods to obtain quantitative results, as will be conducted in Section 5.3. Nevertheless, this section uses simplified equations to highlight important parameters that can affect the variability of the currents in PN junctions as well as the collector current in SiGe HBTs. For brevity, base current variability is not addressed. For clarity, the term "variability" in this work refers interchangeably to the relative standard deviation (RSD), which is the standard deviation divided by the mean. RSD provides a good method for comparing quantities that vary by orders of magnitude (such as the current in a transistor).

5.2.1 Variability in PN Junctions

The forward-bias current in a PN junction can be modeled in the form of $I = I_S(e^{qV/kT}-1)$, where q is the electron charge, k is the Boltzmann constant, T is absolute temperature, I_S is the reverse-saturation current, and V is the applied voltage. As will be shown by

measurement presented in Section 5.4, variations in I exist even at the same V. Therefore, we examine the possible causes of variability in I_S.

The I_S for an abrupt N^+P junction is [97]

$$I_S = \frac{qAD_{nP}n_i^2}{W_{nP}N_A}e^{\Delta E_g/kT} \tag{5.1}$$

where A is the cross-sectional area, D_{nP} is the diffusion constant for electrons in the P region (related to mobility, μ_{nP}, through Einstein's relation $D_{nP} = \mu_{nP}kT/q$), n_i is the intrinsic carrier concentration, N_A is the ionized doping concentration (which can be different from the nominal doping concentration due to incomplete ionization), and ΔE_g is the bandgap change due to either heavy doping or mechanical stress in the P region, as will be explained below. Unless otherwise specified, doping concentration in the present work always refers to the ionized doping concentration. W_{nP} is the length of quasi-neutral P region for short diodes, which is replaced in long diodes by L_{nP} ($\ll W_{nP}$), the diffusion length of minority electrons in the P region.

Area (A) is expected to be constant across temperature and does not contribute to the temperature dependence of variability. However, most other parameters depend on doping concentration and temperature. For example, μ_{nP} depends on doping concentration due to both impurity and phonon scattering [98]. The ionization of dopants (affects N_A) depends on the nominal doping concentration and the lattice temperature [99]. W_{nP} depends on the location of the metallurgical junction, as defined by the doping profile, while L_{nP} is determined by the doping-dependent carrier lifetime [100].

ΔE_g can come from two mechanisms. The first mechanism is bandgap narrowing (BGN) due to heavy doping concentration, as shown in Figure 5.2(a) [101]. Therefore, doping variation leads to ΔE_g variation. To the first order, BGN can be assumed to be independent of temperature [102]. The second mechanism is the bandgap change due to the mechanical stress, as shown in Figure 5.2(b) for the z-axis stress as an example [103]. The mechanical

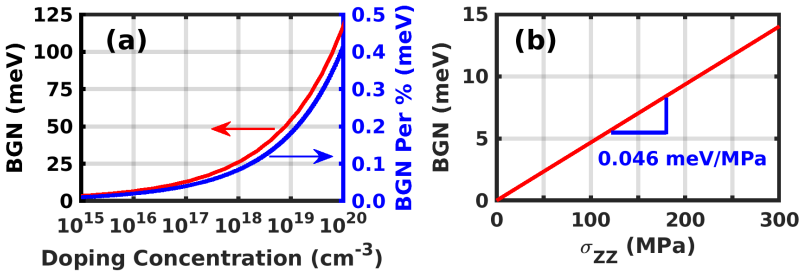

Figure 5.2: (a) Left y-axis: BGN narrowing versus doping. Right y-axis: BGN due to 1% change of doping concentration. (b) BGN due to stress vs the z-axis stress (zero stress in other directions).

stress can come from the back-end-of-line (BEOL) process and the layout placement [104, 105], or oxide isolations commonly used in semiconductor processing [106]. Therefore, variations in process conditions lead to variations in stress level, which leads to variation in ΔE_g.

For both doping- and stress-induced ΔE_g, due to the unique properties of the Arrhenius function for thermal activation (i.e., $e^{-E_a/kT}$), the induced variability of ΔE_g on I_S tends to be enhanced exponentially at decreased temperatures. This is explained in Appendix A and will be referred to many times in this work.

5.2.2 Variability in SiGe HBTs

In SiGe HBTs, the collector reverse-saturation current is given by [35]

$$I_{C0,SiGe} \propto I_{C0,Si} \frac{e^{\Delta E_g(0)/kT}}{1 - e^{-\Delta E_g(grade)/kT}} \tag{5.2}$$

where $I_{C0,Si}$ is in the same form as (Equation 5.1), $\Delta E_g(0)$ is the Ge-induced bandgap narrowing at the EB metallurgical junction, and $\Delta E_g(grade)$ is the difference of bandgap across the quasi-neutral base due to position-dependent Ge grading. As in PN junctions, variability of $I_{C0,Si}$ can be enhanced exponentially at low temperature due to ΔE_g variations.

In addition, $\Delta E_g(0)$ is affected by both the Ge and the doping profile. Variability of

either profile near the EB junction can shift the EB Ge percentage, causing variability in $\Delta E_g(0)$. Similar to the ΔE_g in PN junctions, variability of $\Delta E_g(0)$ tends to amplify the variability of $I_{C0,SiGe}$ exponentially at decreased temperatures.

At low enough temperatures, the collector current in advanced SiGe HBTs (typically with a thin base) is accounted by direct tunneling [54, 76], and the model for collector current is different from (Equation 5.2). The tunneling current is directly proportional to the emitter electron density ($\approx$ emitter doping) multiplied by tunneling probability (P). For simplicity, consider a constant and abrupt base doping profile, which simplifies P to [107]

$$P = \exp\left(\frac{-2W_B}{\hbar}\sqrt{2m_e(E_{C,B} - E_{Fe})}\right) \tag{5.3}$$

where $\hbar$ is the reduced Planck constant, W_B is the quasi-neutral base width, m_e is the effective electron mass, $E_{C,B}$ is the conduction band energy in the base, and E_{Fe} is the electron quasi-Fermi level in the emitter. The main source of variation in this case is W_B and $E_{C,B}$, both of which vary with doping profile.

5.3 TCAD Investigations

Given the previous list of potential mechanisms that can affect the variability in current across temperature, calibrated TCAD simulations have been conducted to examine their quantitative impact. In particular, this section will show that

1. In PN junctions, BGN and stress variations tend to increase the current variability at low temperatures, while other mechanisms have a weaker effect.

2. In SiGe HBTs, the Ge profile amplifies the effects of other sources of variability on the collector current variability at low temperatures.

3. With the presence of tunneling, collector current variability increases with decreasing temperatures if the doping variation is in the emitter or the collector, while it saturates

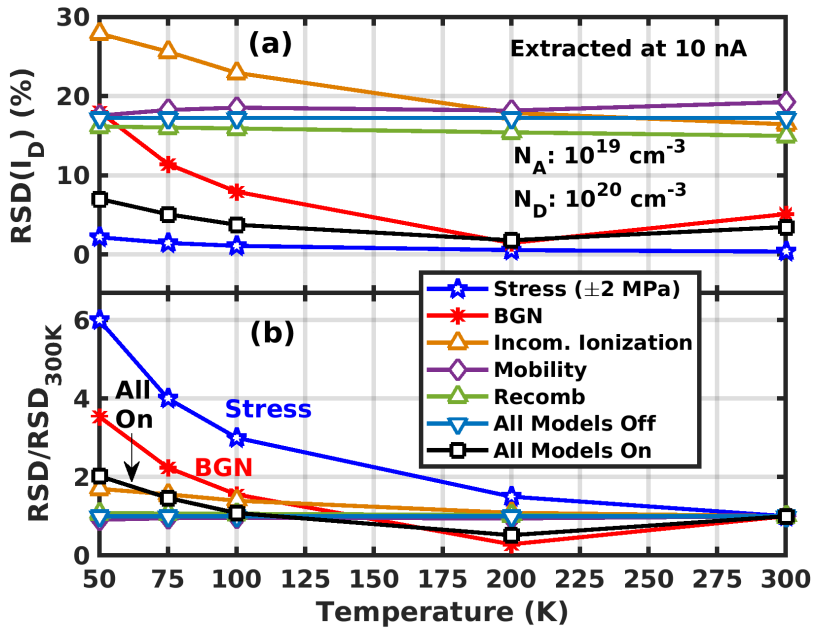

Figure 5.3: (a) Variability of diode current across temperature, with different models individually turned on. (b) Variability in (a) normalized to 300 K values.

at lowering temperatures if the doping variation is in the base.

5.3.1 Variability of PN Junctions

We first use PN junctions to demonstrate the effects of fundamental physical mechanisms on the temperature trends of terminal current variability. Towards this goal, a hypothetical PN junction profile with a P doping of 10^{19} cm^{-3} and an N doping of 10^{20} cm^{-3} was created. The P and N doping profiles are defined as Gaussian decay. The peak P and N doping are then varied by $\pm 20\%$ to create a batch of profiles. To evaluate the effects of physical mechanisms on the variability of this batch, the simulations include only one operative physical mechanism (model) at a time, while other models are turned off or set independent of doping. The models used in this study include: the Philips unified mobility model [98], the SRH recombination model [100], Jain-Roulston bandgap narrowing [101], and incomplete ionization [99]. A separate batch of profiles was created without doping variations, but with varying mechanical stress. Uniform stress of 230±2 MPa applied perpendicular to the transport plane of the device was simulated with a stress-dependent

bandgap model [108]. 230 MPa is an estimate of stress level due to BEOL routing (based on [104]), and 2 MPa variation is feasible given that the typical stress levels due to BEOL or oxide isolation are hundreds of MPa [104, 109]. The stress-dependent bandgap model calculates the stress-induced band edge shift in both the conduction and valence band.

The simulations were conducted across temperatures from 300 K to 50 K (the lowest temperature for robust numerical convergence). At each temperature, the variability of the batch was extracted as the variability of current at a forward bias voltage that corresponds to 10 nA of current for the nominal profile. Outside the non-ideal region in the highest and lowest injection regions, variability is independent of extraction bias because the variability shows up as parallel shifts of current (in log scale) versus voltage. The variability across temperature is plotted in Figure 5.3(a) and the variability normalized to its room temperature value is shown in Figure 5.3(b) to showcase the temperature trends in variability due to each mechanism.

From Figure 5.3(a), it can be seen that BGN or stress alone has a smaller variability than that of the other models. However, Figure 5.3(b) shows that the relative increase of variability for these two mechanisms is quite large. This is due to the presence of variation in the exponent of $e^{1/kT}$, as explained in Section 5.2. In Figure 5.3(b), incomplete ionization also increases the variability at low temperatures, while all other models have a negligible impact on the temperature trends of variability. When all models are turned on (with both doping and stress variation), the variability still increases at low temperatures, but not as much as when stress or BGN is turned on individually. As will be explained, the effect on variability due to multiple mechanisms is not necessarily additive, but can instead be dominated by the strongest factor. These simulation results show that bandgap-related mechanisms, such as BGN or stress, have a strong effect in increasing the variability at low temperatures, while other mechanisms have a weaker impact.

However, this is not the complete picture, because the variability due to each mechanism is also dependent on the nominal doping concentration. To obtain a more complete

picture, we examine the ratio of variability between 50 K and 300 K for each mechanism across all combinations of P and N doping levels between 10^{16} cm^{-3} and 10^{21} cm^{-3}. The results are shown in Figure 5.4.

A few observations can be made. First, based on the symmetry across $N_D = N_A$ (donor and acceptor doping), the variability of current from minority hole versus minority electron is similar. Therefore, swapping P and N doping species in a junction results in a negligible change in its variability, even though the material parameters are slightly different. Second, in terms of the range of doping where each effect is pronounced, BGN results in larger low-temperature variability only if both dopings are larger than 10^{17} cm^{-3} (near the onset of BGN), with a nearly 10x increase at high doping. This can be understood because the slope of BGN versus doping increases with doping concentration, as shown in Figure 5.2(a). The effect of stress variation is independent of doping concentration. Even though the bandgap change due to BGN or stress is independent of temperature, their effect on the current variability is amplified at low temperature (explained in Section 5.2 and Appendix A). Incomplete ionization only has a slight effect around 5×10^{18} cm^{-3} (where the dopant ionization strongly depends on nominal doping). Mobility only increases the variability if the doping is below 5×10^{17} cm^{-3}, because at higher doping levels the mobility begins to saturate with a further increase in doping (i.e., becomes less sensitive to doping) [98]. Recombination produces a small effect in general because the SRH recombination current is negligible compared to the diffusion current. Overall, we see that for ranges of doping concentration typical in modern devices ($> 10^{17}$ cm^{-3}), BGN and stress can have a strong effect on the increase of variability at cryogenic temperatures, while the effects due to other mechanisms are weaker.

So far, individual mechanisms are examined one at a time. In real devices, all physical mechanisms affect the variability simultaneously. Figure 5.5 shows simple scenarios when only two mechanisms are combined. The first scenario, shown in the first row of Figure 5.5, is when the effect of one mechanism dominates that of the other. In this case,

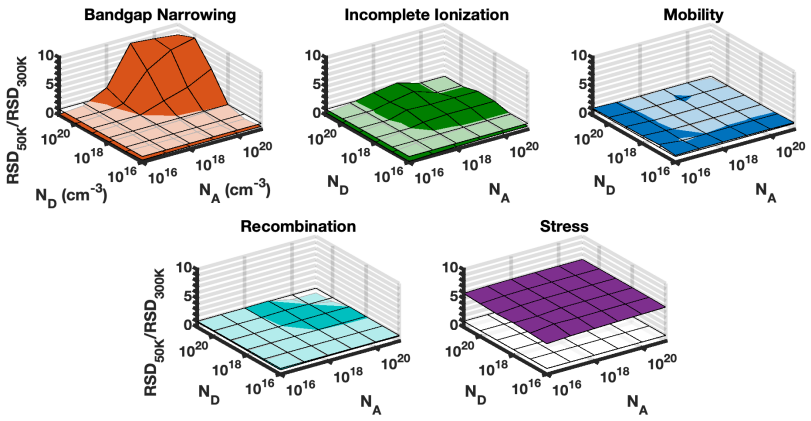

Figure 5.4: Effect of various mechanisms on diode current variability across doping concentration.

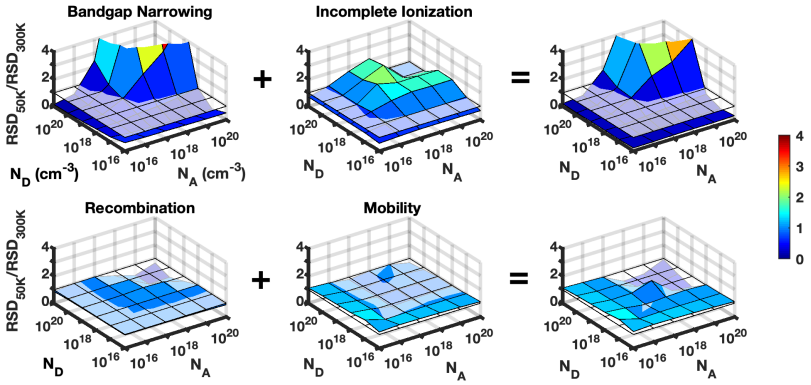

Figure 5.5: Illustration of how two physical mechanisms combine to generate the resulting variability for a PN junction.

BGN has a stronger effect in enhancing the variability than incomplete ionization (refer back to Figure 5.4), and therefore the variability with both mechanisms is similar to the variability when only BGN is present. The second scenario is when two effects are similar in strength, as shown in the second row of Figure 5.5. Here, two effects are positively combined when both mechanisms are present. However, with more than two mechanisms present, the combined effect is not necessarily the addition of individual effects, but can be subtractive (e.g. Figure 5.3). This can be simply understood that variations in the numerator and the denominator of an equation cancel each other and result in less or no overall variation. In this case, numerical simulations are needed for quantitative results. Overall, it is still important to distinguish the significant mechanism in the operation of a device,

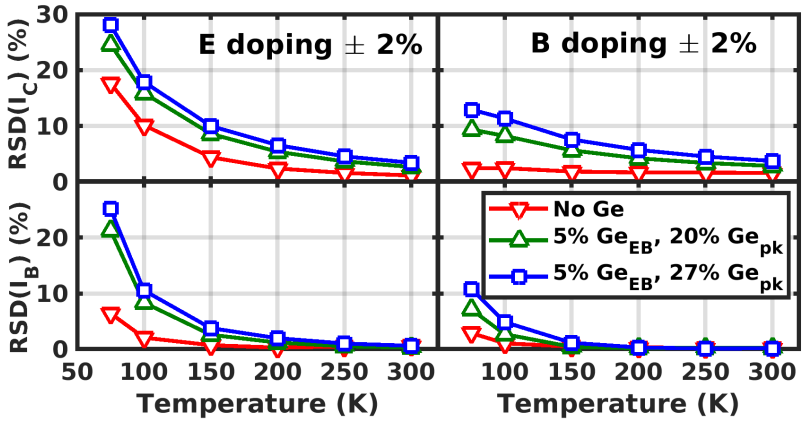

Figure 5.6: Effects of Ge percentage on the variability of current.

because they set the main trend of variability across temperatures.

5.3.2 SiGe HBT Variability

In addition to the mechanisms presented for PN junctions, the over-temperature variability of SiGe HBTs is also affected by the presence of the Ge profile and the tunneling transport mechanism. To study their impact, a calibrated SiGe HBT profile from our previous work has been used [2]. The peak doping concentration and the Gaussian standard deviation in the base and emitter are varied by $\pm 2\%$ from their nominal values. For baseline comparisons, a reference BJT doping profile was created by setting the Ge content in the SiGe HBT to zero. BGN, mobility, recombination, and a non-local direct tunneling model [110] are included in the simulations. Based on the knowledge of the epi growth technique in this process technology, the Ge diffusion is assumed to be negligible compared to the diffusion of other doping species, resulting in negligible Ge profile variability compared to doping variation. However, the EB Ge fraction still depends on the EB junction location, which will be changed by the introduction of doping profile variations.

Figure 5.6 highlights the variability of base current (I_B) and collector current (I_C) across temperature for different Ge profiles. With no Ge (i.e., a BJT), there is a slight increase in variability at low temperatures when the variation is located in the base or emitter region. The base and emitter doping are high enough such that their variations also

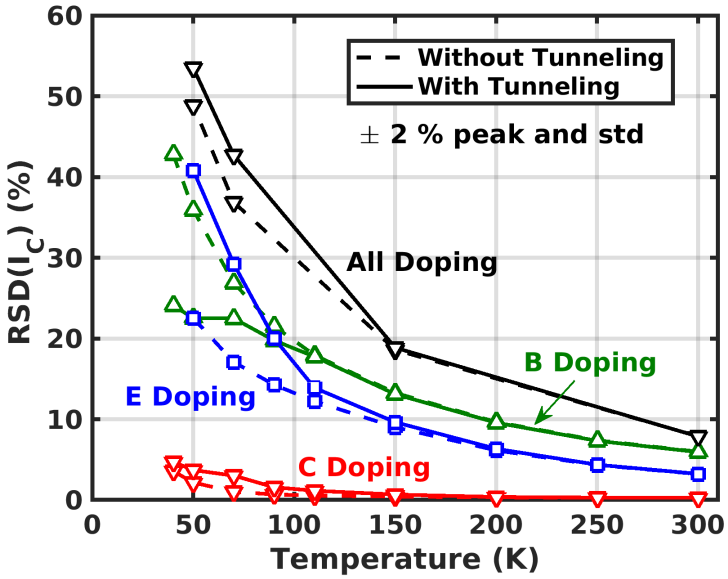

Figure 5.7: Effects of tunneling on the collector current variability across temperature, combined with doping variation in the emitter, base, or collector regions.

change the bandgap due to BGN, whose variations are exponentially amplified at low temperatures. The addition of Ge universally enhances the variability even more than for the BJT, for both I_C and I_B. This is expected due to the variation of EB Ge fraction, as explained in Section 5.2. The collector doping variation (not shown here) results in a much smaller change for all Ge profiles and temperatures. Through these results, it has been confirmed that the presence of Ge, by itself, will increase the variability of SiGe HBTs at low temperatures compared to a similarly-constructed Si BJT.

Next, we investigate the effects of tunneling on over-temperature variability. In Figure 5.7, doping variations in the base, emitter, or collector profiles are simulated, both with and without the tunneling mechanism turned on. As can be seen, if the doping variation is in the emitter or collector region, tunneling increases the variability. In contrast, if the doping variation is in the base, the variation saturates at low temperatures with tunneling. This is because emitter and collector doping only affect W_B and not $E_{C,B}$, while base doping affects W_B and $E_{C,B}$ in (Equation 5.3) in opposite ways, thereby partially canceling its impact on current variability. The difference in variability versus temperature (increasing

versus saturating) can provide insight into the precise region that has the dominant doping variation.

5.4 Measurement Results

This section compares TCAD simulation with measurement to provide more insights into the dominant factors for over-temperature variability in real devices.

Measurements were obtained from a batch of PN junctions and SiGe HBTs. All devices are from GlobalFoundries 130-nm BiCMOS technologies [111]. A total of 13 high-performance (HP) NPN SiGe HBTs (on the same die) and 24 N-well substrate diodes (12 from each die) were measured across temperature, from 300 K to 15 K. The dies come from the same multi-project wafer run. All SiGe HBTs have the exact same layout from the device to probe pad, while the diodes have a different BEOL wiring. The emitter geometry of SiGe HBT is 0.12x9 $\mu\mathrm{m}^2$. The cross-sectional area for the diode is not well defined due to the lateral ring-diode construction. The die size is 5x6 mm^2, while the SiGe HBT structures only occupy 0.9x2.4 mm^2. The measurements were performed in a closed-cycle probe station capable of reaching a temperature of 12 K. An Agilent 4156C Semiconductor Parameter Analyzer was used for device characterization. The sample plate temperature was controlled to better than 0.1 K using calibrated temperature sensors. The measurement setup, including temperature fluctuations, probe contact, and the resolution of 4156C Analyzer, was verified to introduce negligible variability to the measured data.

5.4.1 PN Junction Diodes

The I-V curves of the diodes and the variability of the current extracted at a fixed voltage are shown in Figure 5.8. Additionally, a TCAD model calibrated to the diode doping was simulated, and is shown in Figure 5.8 using a dashed line. As shown, the variability of the diode current increases at low temperatures. Since the peak doping of the diode is around $N_A = 10^{15}\ \mathrm{cm}^{-3}$ and $N_D = 10^{18}\ \mathrm{cm}^{-3}$, only stress and mobility variations can explain the

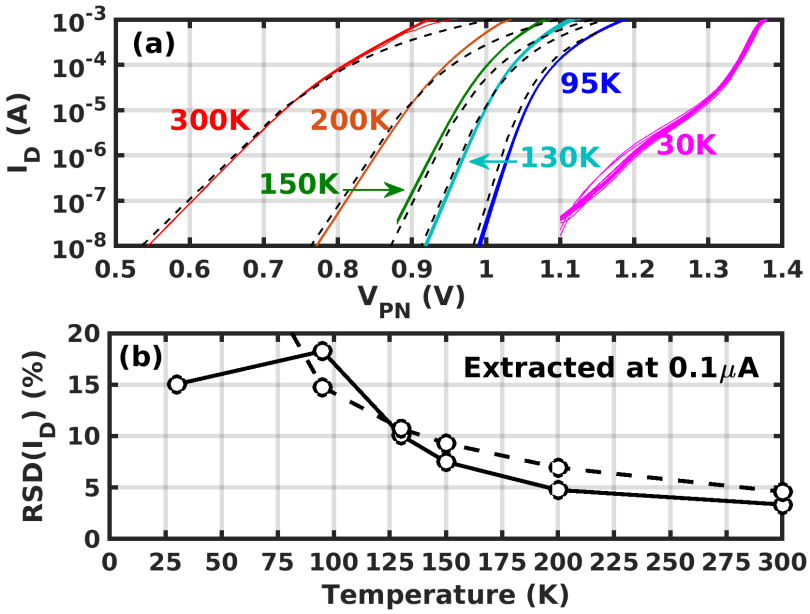

Figure 5.8: (a) I-V curve of PN diodes across temperature. (b) Variation of diode current versus the temperature. Extracted at the same voltage for the whole batch, where the voltage is selected at each temperature that corresponds to 0.1 μA for one of the diodes. Solid: measurement. Dash: TCAD simulations.

increase of variability at low temperatures, based on Figure 5.4. However, the increase of variability at low temperatures due to mobility variations produce a far lower value, even for a 30% doping variation, which is considered unrealistically large for two neighboring devices on the same die. In the end, it was found that a 20 MPa standard deviation of stress with mean stress of 230 MPa (uniform and perpendicular to the die surface) is needed to model the observed increasing trend of variability. Such variations are plausible because the BEOL wiring stacks are different between the measured diodes. Based on the stress value from the BEOL given in [104, 106], the presence of one additional metal layer above the device can change the stress by 100 MPa. The average stress is found to have a negligible impact on variability. Instead, the standard deviation alone determines the value and the temperature-dependent trend of the variability.

Overall, we find that even simple PN junctions can exhibit increasing variability at low temperatures when mechanical stress variation is applied to the device.

5.4.2 SiGe HBTs

Next, we examine the variability of SiGe HBTs, which are more complicated than diodes due to the presence of heavy doping, Ge profile, and tunneling processes. The Gummel characteristics of the measured SiGe HBTs are shown in Figure 5.9(a), while the variability of the collector current (I_C) is shown in Figure 5.9(b). The curves at 30 and 15 K largely overlap due to the presence of tunneling.

First, we observe that the variability of I_C increases at low temperatures, consistent with our preliminary work [3]. To gain additional insight, a SiGe HBT calibrated to 300 K I_B and I_C and over-temperature I_C was simulated and is shown in dashed lines in Figure 5.9(a). The doping and Ge profile were calibrated to produce the correct Gummel curves at room temperature, while the amount of doping and stress variation was calibrated to reproduce the increasing variability across temperature. Once this calibrated set of profiles is obtained (referred to here as the reference), we selectively turned off various physical models to assess the impact of each operative mechanism.

When the bandgap narrowing is turned off, the variability increases faster than the reference until 78 K, below which the variability begins to saturate. In addition, without tunneling, the variability is larger below 75 K compared to the reference. Both facts consistently suggest, based on Figure 5.7, that the base doping variation plays a stronger role than the emitter doping variation to produce the observed temperature trends in variability. Incomplete ionization is found to have a negligible effect in this SiGe HBT because the base and emitter doping are far above 10^{18} cm^{-3}, where the semiconductor is degenerate and the impurity band overlaps with the valance or conduction band. SRH recombination also has a negligible effect due to the thin base nature of the device.

Unlike the measured diodes, all of the SiGe HBTs have the same drawn layout. However, mechanical stress is found, surprisingly, to also play a critical role in reproducing the increase in variability at low temperatures. Without stress, the variability can be tuned to increase at low temperatures, but the variability at room temperature also increases substan-

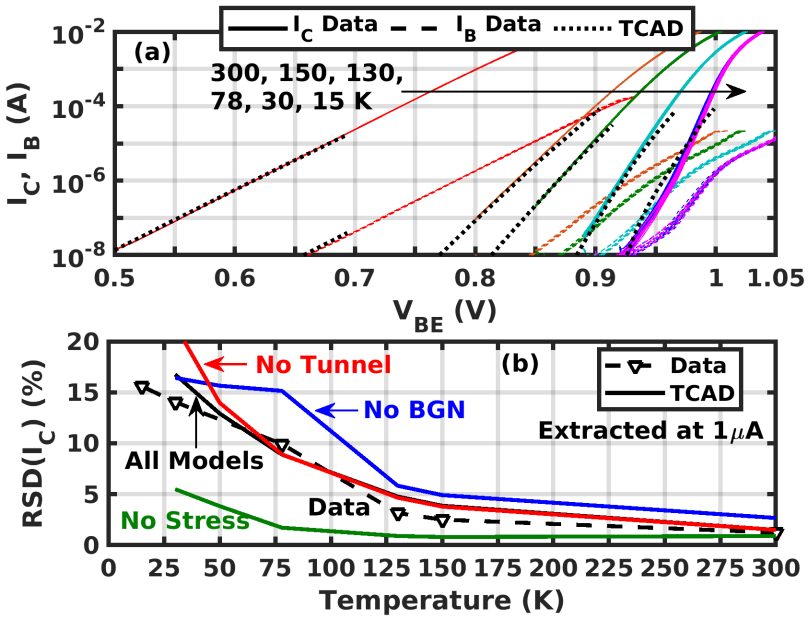

Figure 5.9: (a) Gummel characteristics of SiGe HBTs across temperature. Solid: measurement. dash: TCAD simulations. (b) Variability of collector current across temperature. Solid: TCAD simulations. dash: measurements.

tially with more doping variation (needed to produce a larger low-temperature variability), which deviates from the measured data. In the end, a stress variation of 13 MPa in the z-axis is included in the simulations. Since there are 5 lower metal layers directly above the device, and two higher metal layers within 2 μm of the contacts to the device, the mean stress is very likely above 400 MPa based on [104]. Therefore, 13 MPa corresponds to a 3-6% of stress variation within a die. This is plausible, given the presence of randomized foundry patterned fill, the potential stress from another nearby layout, or variations of shallow trench isolation (STI) [109]. As a sanity check, separate simulations (not shown) reveal that, even if the Ge profile has a finite but small variability (as it inevitably will have in reality), the resulting variability is much smaller compared to the effects of doping or stress variations.

In summary, it is found that BGN, tunneling, and mechanical stress all play critical roles in predicting the observed temperature trends of collector current variability in SiGe HBTs.

5.5 Discussion

Given the picture this far, it is of interest to explore potential methods for mitigating variability at low temperatures.

Among the dominant mechanisms, doping-induced BGN variability can be reduced by tighter control on doping profiles. Mechanical stress can be mitigated by keeping the layout identical for each device. However, as shown by the present work, even an identical layout can create a few percent of variability across the die. Therefore, for key circuits, common-centroid layouts should be used to create a best-case matched pair. However, common-centroid layouts can create routing difficulties if the minimization of BEOL stress is also required.

In SiGe HBTs, reducing I_C variability requires a Ge profile that is constant near the EB junction, namely, by using a Ge ledge [35]. However, a tradeoff exists between the manufacturability and stability of Ge ledge (now protruding further into emitter), versus the reduced variability across temperature. For short-base devices where the tunneling is present, mitigation of I_C variability should focus on minimizing the variability of the emitter and collector profiles. Although base profile variability has less impact on I_C, it still needs to be minimized to prevent excessive variability in I_B and current gain. Overall, detailed knowledge about a process technology is required to determine which one of these techniques will offer the most cost-effective mitigation strategy.

5.6 Summary

This study examines various physical mechanisms that can cause temperature-dependent variability in the terminal current of PN junctions and SiGe HBTs. For PN junctions, the main cause of increased variability at low temperatures is bandgap narrowing due to heavy doping as well as the bandgap change induced by mechanical stress. In SiGe HBTs, the mere presence of Ge further increases the variability at low temperatures, while tunneling

can increase the variability depending on the region where the doping variation occurs. Other mechanisms, such as incomplete ionization, mobility, and recombination, play only minor roles in determining the variability at cryogenic temperatures.

CHAPTER 6

COMPACT MODELS OF SIGE HBTS AT CRYOGENIC TEMPERATURES

To have successful circuit designs for quantum science, a compact model for SiGe HBTs is required. Existing models offer limited infrastructure for further development, and a new foundation is desired. This chapter finds that the HICUM/L0 compact model can be successfully applied to SiGe HBTs operating at cryogenic temperature, verified at both device and circuit level. HICUM/L0 provides not only the DC and small-signal modeling, but also large-signal modeling that is needed for circuit blocks such as mixers. The development of the compact model in this chapter should pave the road for future circuit design using SiGe HBTs.

6.1 Introduction

Practical implementations of quantum computers (QC) require large-scale integrated electronics operating at temperatures as low as tens of mK [13]. Among the various possible candidates for such electronics, SiGe HBTs provide unique advantages. Compared to CMOS at fixed technology node, SiGe HBTs offer lower noise, higher bandwidth, and lower power dissipation. Compared to III-V technologies, SiGe HBTs provide comparable performance but maintain attractive CMOS integration compatibility. Additionally, proposed schemes for control and readout electronics for quantum computers require circuit blocks such as low-noise amplifiers (LNAs), mixers, RF switches, and microwave signal generators [13, 112, 113]. These circuits are routinely implemented in SiGe technology with impressive performance results. This fact, combined with the proven performance of SiGe HBTs operating at deep cryogenic temperatures, makes a compelling case for using SiGe HBTs for emerging quantum computing control and readout circuits [114]. Presently missing, however, is a viable SiGe HBT compact model for such applications. A publicly-

available, yet accurate compact model, requiring only simple extraction routines, is thus highly desirable. This study demonstrates the HICUM/L0 (V1.32) compact model for advanced SiGe HBTs operating at 12 K [115].

Among the compact models in the literature for cryogenic SiGe HBTs, few provide the comprehensive features for QC application needs. Numerous small-signal equivalent circuits have been proposed, where one set of parameters is extracted for each bias point [116, 117]. However, this approach limits the circuit design to bias points with known parameters. More fundamentally, small-signal models require an elaborate expansion in order to simulate large-signal behavior, such as gain compression, intermodulation, and frequency conversion, which are key in signal generation and up/down-conversion blocks in QC control electronics. Other compact models in the literature either model the device only for DC [118, 119], or at higher temperature ranges [120]. Though they show good precision across a wide range of temperatures, the most compelling case for using SiGe HBTs centers around RF circuits, making it a severe constraint to have a DC-only model. Exotic modeling approaches, such as neural network-based models, lack physical basis and are prone to serious error when the bias point lies outside the dataset range [121]. In addition, an often ignored aspect is the availability of models in commercial circuit simulators. The implementation and verification of a compact model is both challenging and time-consuming. HICUM/L0 is a standard model for SiGe HBTs and is available in mainstream commercial circuit simulators, lowering the barrier for adopting SiGe HBTs for emerging QC applications. Overall, it is highly desirable for the community to have a standard compact model capable of DC, small-signal, and large-signal modeling. This study demonstrates such a model.

As proof of efficacy, the HICUM/L0 compact model is first extracted for an advanced SiGe HBT and then applied to model a wideband (1–17 GHz) RF amplifier operating at 12 K. A slightly modified parameter extraction routine is proposed. This SiGe HBT model can predict small and large-signal operations over the range of biases relevant for RF cir-

cuit design, thus supporting emerging QC needs. This is the first demonstration of a DC, small-signal, and large-signal compact model for SiGe HBTs operating at deep cryogenic temperatures.

6.2 Device Technology and Measurement Setup

The SiGe HBTs and RF circuits in the present work are from the GlobalFoundries SiGe BiCMOS-8XP technology platform with f_T/f_{MAX} of 260/320 GHz at room temperature [111]. A 0.12x8 μm^2 SiGe HBT with common-emitter CBEBC layout, and the associated open and short de-embedding structures have been characterized. DC and RF characterizations were performed at 12 K using a closed-cycle RF probe station. S-parameters and P1dB (input-referred 1-dB gain compression point) measurements from 0.1–25 GHz were obtained using an Agilent E8363B Network Analyzer. SOLT calibration was performed at 12 K before S-parameter measurements. Noise figure was measured using a N9030A PXA and a N4002A noise source. DC characterization was obtained using an Agilent 4156C Semiconductor Parameter Analyzer.

6.3 Device Modeling Results

Sample modeling results are shown in Figure 6.1 and Figure 6.2 for DC, and in Figure 6.3 for RF characteristics. The agreement between simulation and measurement is satisfactory. The cryogenic parameter extraction routine is the same as that reported in [122, 123], except for the assumption of device temperature, as will be explained below. The model parameters can be extracted and optimized in simulators like Advanced Design System (ADS). Alternatively, the HICUM group is developing an open-license compact modeling toolkit (DMT) that supports the full model extraction [124].

One modification to the extraction routine is to assume the device temperature to be 50–60 K during the extraction of saturation currents. This is an unphysical, yet necessary workaround for the non-ideal temperature trend of the collector current in advanced SiGe

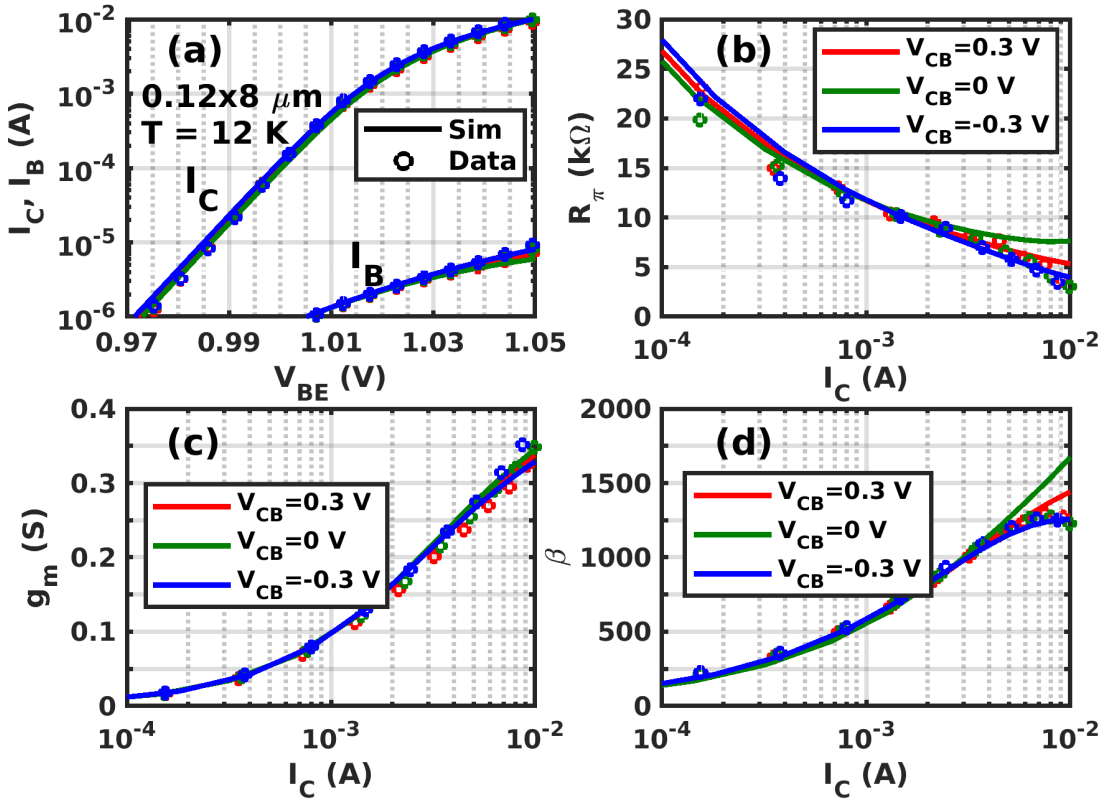

Figure 6.1: Simulated and measured (a) Gummel characteristics, (b) DC base resistance ($R_\pi = \partial V_{BE}/\partial I_B$), (c) DC transconductance ($g_m = \partial I_C/\partial V_{BE}$), (d) DC current gain ($\beta = I_C/I_B$) at 12 K.

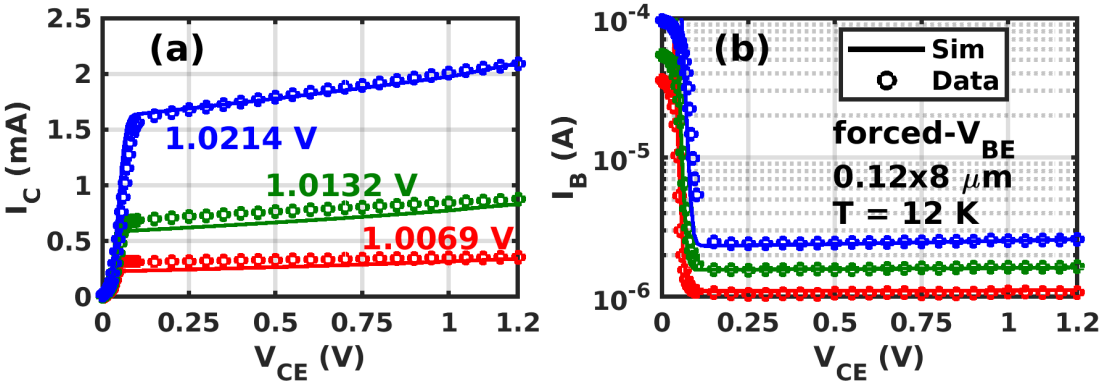

Figure 6.2: Simulated and measured output curves by forced V_{BE}: (a) collector current and (b) base current.

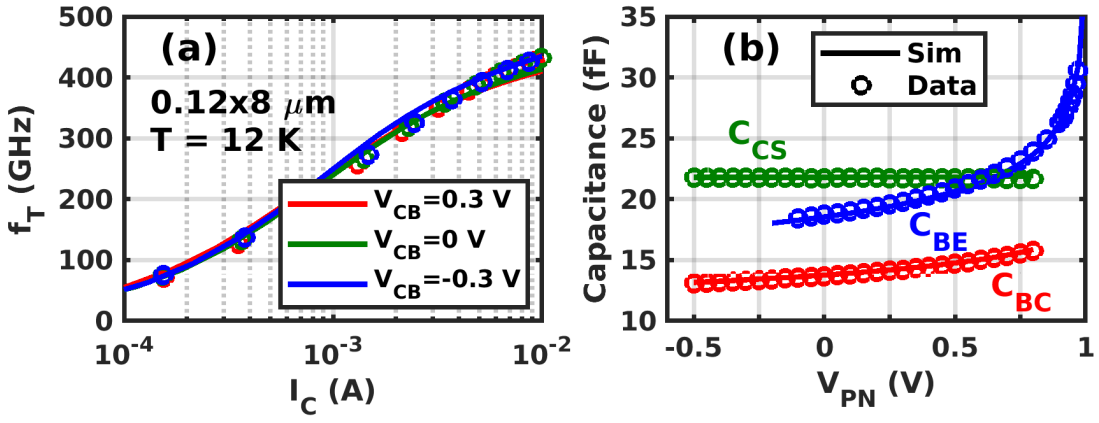

Figure 6.3: Simulated and measured (a) transit frequency and (b) junction capacitance at 12 K.

HBTs [76, 65]. Below 50 K, the slope of collector current saturates to a constant value determined by the direct tunneling, which is not captured by the drift-diffusion formulation in the HICUM model. If the ambient temperature of 12 K is used for extraction and simulation, two issues arise. First, the slope of collector current is far from the measurement (by 5–6 times) and needs a large ideality factor (5 to 6) to compensate. Second, the small magnitude of the thermal voltage V_T ($= k_BT/q$) creates convergence problems in simulators. A workaround is to set the device temperature to around 60 K in the simulator in order to match the slope of collector current (I_C) to measurement with ideality factor fixed to unity. For circuit simulations discussed in Section 6.4, the temperature of other (non-SiGe HBT) components in the circuit are left at 12 K.

Setting an artificially higher device temperature leads to higher thermal noise generated by the series resistances inside a device. However, since advanced SiGe HBTs have small series resistance ($\ll 50\ \Omega$), their thermal noise contribution at a typical bias level of 0.5–5.0 mA of I_C is negligible. A quick estimate of the $4k_BT$ thermal noise from an extremely high value of $50\ \Omega$ series base resistance gives 1 K versus 0.17 K noise temperature for an ambient temperature of 60 K versus 10 K. This error can be significant for LNAs that have only a few kelvins of noise temperature, and requires more sophisticated compact model (presently under development). However, for other RF circuits, such a small difference does not pose a serious concern, especially given the ease of model implementations and

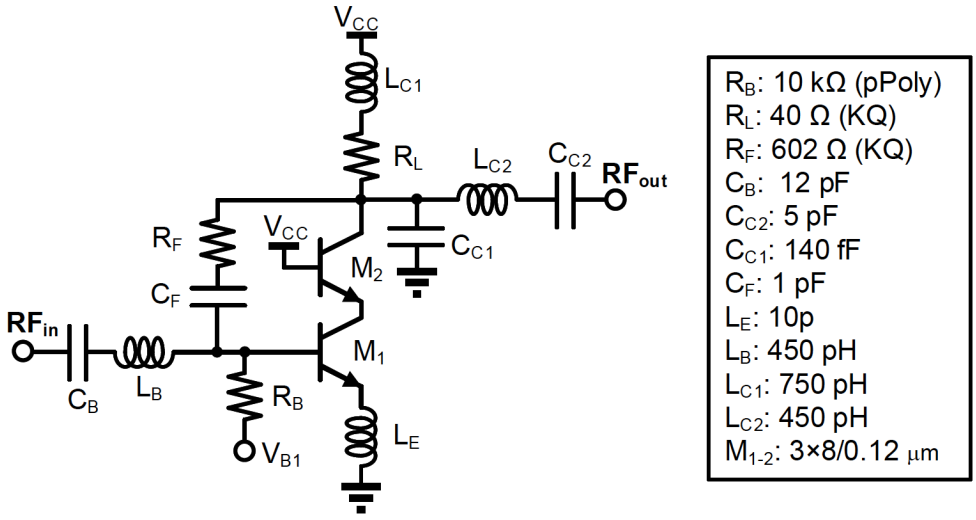

Figure 6.4: Schematic diagram for the 1–17 GHz SiGe LNA at 12 K.

fast convergence with higher device temperature.

Overall, the model shows good accuracy and convergence speed for most range of collector current between 0.1–10 mA, where most RF circuits will be biased.

6.4 Circuit Verification

In addition to accurate model at the individual device level, an important goal for any compact model is its predictive capability at the circuit level. In particular, QC applications demand circuits operating at 4–8 GHz for superconducting qubits and potentially up to mm-wave frequencies for future, higher temperature qubits. To verify that HICUM/L0 model can predict circuit behavior at cryogenic temperatures for both current and future needs, the S-parameters, P1dB, and gain at P1dB of an LNA were measured and compared to simulations at 12 K. The LNA is a wide-band (1–17 GHz) amplifier that utilizes a resistive feedback topology with a cascode amplifier core, as shown in Figure 6.4.

A major advantage of having a robust, cryo-T DC and small-signal compact model lies in the ability to sweep the bias conditions during the design phase, and thereby immediately see the resulting changes at the circuit-level. This gives circuit designers the freedom to optimize the bias point very quickly and accurately. To evaluate the present compact

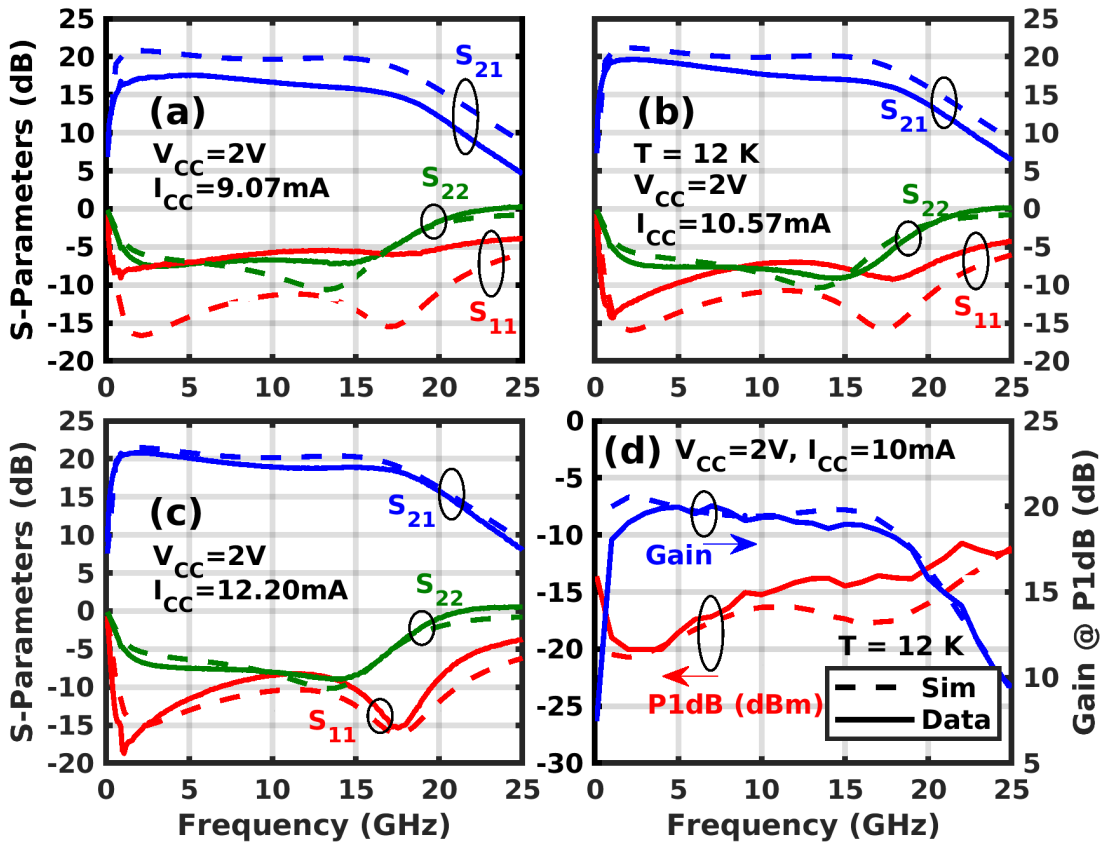

Figure 6.5: (a-c): S-parameter simulations and measurements, (d) P1dB and power gain at P1dB for the 1–17 GHz LNA.

model for this purpose, the LNA was measured under multiple bias conditions and compared with simulations. The simulations include layout parasitics from RC extraction and the electromagnetic (EM) simulations for the inductors and signal traces. For EM simulations, the metal conductivity was increased by 10x to model the behavior of aluminum and copper with cooling over this temperature range, while the dielectric constant was kept the same as at room temperature due to its expected weak temperature dependence. Values of intentional resistors were scaled by a correction factor based on a separate characterization of resistors at room temperature and at 12 K (not reported here).

In Figure 6.5(a)-(c), simulated S-parameters of the LNA at three bias conditions are compared to measurements. Overall, the model matches the measurement data reasonably well across both frequency and bias. It can be noticed that higher bias shows closer agreement to measurement for gain (S_{21}) and input matching (S_{11}). To gain further insight on

the source of discrepancy between the simulation and measurements, sensitivity analysis was performed for S-parameters with respect to parameters at both the device (within the transistor model) and circuit levels. Parameters at these two levels are plotted in different colors in Figure 6.6. Sensitivity is defined as the change of response (S-parameters) for a 1% change in a given parameter.

As can be seen, circuit components have significant impact on the gain, while device parameters strongly affect the output matching. Input matching is affected by both device and circuit parameters. In particular, as shown in Figure 6.6(a), input matching is strongly affected by emitter and base inductors at the circuit level, and c_{jci0} and r_{bx} (zero-bias internal BC depletion capacitance and external base series resistance) at the device level. For S_{21}, the offset in Figure 6.5(a) may come from possible inaccuracies in the load impedance modeling. Particularly, the temperature coefficient of the resistors embedded in circuit can differ from that of the separately characterized resistors. The coefficient, r_{Ratio}, which characterizes the ratio of resistance at 300 K versus at 12 K for R_L and R_F, is also shown to have a large effect on S_{21} in Figure 6.6(b). In addition, the inaccuracy in the EM simulations of emitter, base, and collector inductors can all contribute to the gain and input matching errors. Interestingly, room temperature simulations (not shown) using an accurate PDK model of the SiGe HBT with EM-simulated passives, also show gain offset compared to measurements, suggesting that some aspects of layout may be systematically missing in the EM simulations. At the device level, c_{js0} (zero-bias SC depletion capacitance), c_{jci0} (zero-bias BC depletion capacitance), and r_{cx} (external collector series resistance) strongly affect the gain roll-off. Enhanced variability of electrical characteristics at cryogenic temperatures can also cause some deviations between the measured SiGe HBT and the SiGe HBTs embedded in circuit [3].

Based on the sensitivity analysis, despite the emphasis of the present study on accurate device models, cryogenic circuit designs also need to separately verify the accuracy of EM simulations and the statistics of resistors at target temperatures to ensure overall simulation

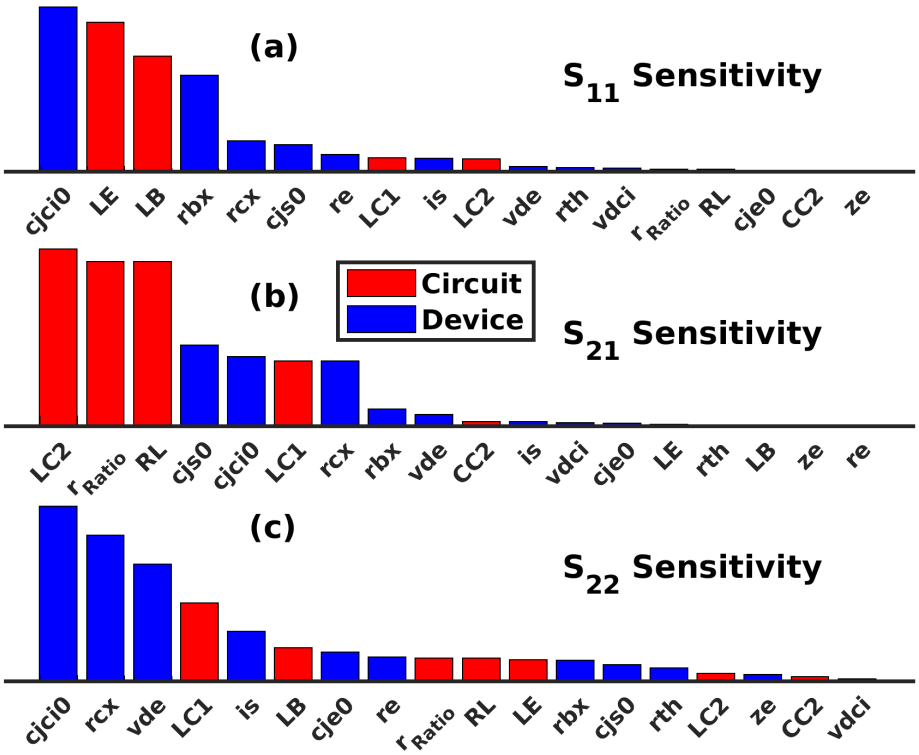

Figure 6.6: Sensitivity of device and circuit parameters that affect the S-parameters. Height is shown in linear scale.

accuracy at the circuit level, which is beyond the scope of this work. Potential inaccuracy in passive modeling aside, the demonstrated model is still promising for QC applications.

A key advantage of a full compact model like HICUM/L0 lies in its ability to simulate large-signal behavior. P1dB of the wideband LNA is used here to demonstrate large-signal operation. In Figure 6.5(d), simulations of P1dB and power gain at P1dB are compared with measurements. P1dB is affected by not only the small-signal gain but also interactions of high-order harmonics generated by the transistors. As can be seen in Figure 6.5(d), the P1dB simulations are close to the measured values, and the gain is well-predicted. At higher frequencies, the discrepancy seen in P1dB vs. simulations may come from the actual load resistance being different from the nominal value due to inherent process variations. In summary, the ability to generate the large-signal and small-signal response with the same model demonstrates a major advantage of this large-signal SiGe HBT cryogenic

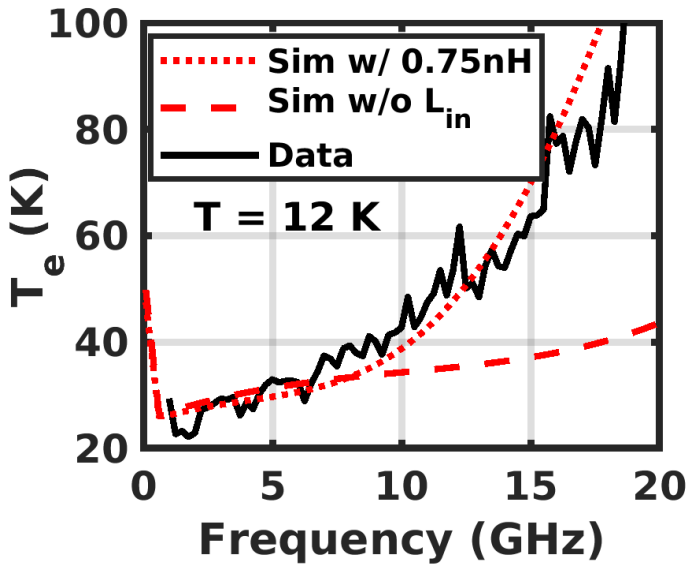

Figure 6.7: Noise temperature simulation and measurement at V_{CC}=2 V, I_{CC} = 10 mA.

temperature compact model over other customized small-signal models.

Finally, the noise performance was simulated and checked against measurements obtained using the cold-attenuator method. As shown in Figure 6.7, the noise behavior can be accurately predicted below 10 GHz. Above that, the wirebond between the cold attenuator and the circuit introduces additional noise, but can be modeled if a reasonable wirebond inductance of 0.75 nH is assumed.

Overall, the HICUM/L0 is shown to accurately model DC, small-signal, large-signal, and noise behavior at an ambient temperature of 12 K across bias conditions relevant for RF circuit designs. Given that the SiGe HBT DC characteristics do not change substantially from 12 K to 70 mK, we expect the model to be valid for mK temperature circuit design. Progress is currently underway to extend the model to a larger temperature range with more accurate extraction and optimization routines. A comparison of the features of the current model with those of other existing compact models for SiGe HBTs at cryogenic temperatures is shown in Table 6.1.

Table 6.1: Comparison of SiGe HBT Cryogenic Compact Models

Ref	Reported Temp (K)	DC	Small Signal	Large Signal	Included in Simulators
[116, 117]	15	x	x		
[119]	43	x			
[118]	93	x			
[120]	43	x	x		
[121]	4.3	x			
This work	12	x	x	x	x

6.5 Summary

This study demonstrates that HICUM/L0 can be successfully applied to SiGe HBTs operating at a deep cryogenic temperature of 12 K, thereby enabling robust circuit design for emerging quantum systems. Compared to other existing models, this model provides DC, small-signal and large-signal modeling at 12 K, while preserving the simple parameter extraction procedures demanded by HICUM/L0 model, and the wide availability as an industry standard in commercial simulators.

CHAPTER 7

THE EFFECT OF DEVICE VARIABILITY AT CIRCUIT LEVEL

Given the variability results from Chapter 4 and Chapter 5, and the compact models presented in Chapter 6, this chapter combines them to study the effect of variability in the circuit level. The first portion of the study is an experimental study that clarifies and confirms the theoretical analysis presented in Chapter 5. The second portion of this study uses a calibrated compact model to study the effect of variability on RF amplifiers that are important for quantum science applications.

7.1 Introduction

Quantum computing has attracted interest from the cryogenic electronics community due to its demands for cryogenic integrated circuits that control and readout qubits [13]. SiGe BiCMOS technology has unique advantages to serve this demand [39]. SiGe HBTs in BiCMOS technology have proven analog performance while maintaining low power consumption [62], and can achieve both low-power and low-noise cryogenic radio frequency (RF) amplifiers [50], both of which are critical for high-fidelity qubit readout. At the same time, CMOS within the BiCMOS platform enables dense digital circuits needed for the qubit control.

A practical issue not addressed thus far in the literature is the fundamental issue of accurate circuit biasing in the presence of increasing device variability at cryogenic temperatures, which is now known to exist in both cooled SiGe HBTs and CMOS[3, 92] technology. As will be shown, variability can cause the mirror ratio of the current mirror (CM) to deviate from its nominal design point, leading to degraded resolution in data converters and unexpected shifts of performance characteristics for RF circuits [93]. Therefore, it is desirable to comprehensively investigate how SiGe BiCMOS CMs behave at cryogenic

temperatures.

Towards this goal, the present work presents measurements of the performance of a variety of CMs implemented in an advanced 90-nm SiGe BiCMOS technology, and compares results at 300 K and 19 K. Depending on transistor types, operating conditions, and other factors, mirror ratios can deviate significantly from their nominal values at 19 K, resulting in CM biasing errors. Based on these findings, best practices for robust CM design for cryogenic operation are summarized. In addition, the causes of CM errors are investigated.

This study is organized as follows. Section 7.2 describes the test structures and the measurement setup. Section 7.3 investigates the CM errors in the presence of bias condition mismatch. Section 7.4 examines the CM errors without bias condition mismatch, which is thus fundamentally limited by process variations and layout effects. Section 7.5 uses a simulated cryogenic SiGe amplifier to showcase the effects of increasing CM errors on circuit-level performance.

7.2 Test Structures and Measurement Setup

All CM test structures were implemented in GlobalFoundries BiCMOS 9HP technology [80]. The platform includes SiGe HBTs with f_T/f_{MAX} of 310/350 GHz, and 90-nm CMOS devices. The variants of CMOS devices include those with regular threshold voltage (V_T = 440/340 mV for nMOS/pMOS) and high-V_T (530/480 mV) for 1.2 V operations, as well as input/output (I/O) devices for 1.8/2.5/3.3 V operation. The present study uses regular-V_T devices in CMs. Die were placed in a closed-cycle cryogenic probe station and measured with an Agilent 4156C Semiconductor Parameter Analyzer. The minimum stable temperature in the cryostat was 19 K, and the sample stage temperature was controlled to better than 0.1 K accuracy using Lakeshore DT-670 temperature sensors.

While there exists a large number of design possibilities, 10 different CMs have been included in the present work, in order to highlight their differences, and hence illuminate best practices. The details of the measured CMs are given in Table 7.1. Throughout the present

Table 7.1: Details of the Measured Current Mirrors

	Type	Topology	In/Out Size (μm)	Ratio
(a)	SiGe HBT	Simple	0.1x1.25x1	1:1
(b)	SiGe HBT	Simple	0.1x10x1	1:1
(c)	SiGe HBT	CC	0.1x1.25x8	1:1
(d)	nMOS	CC	0.5x2x2	1:1
(e)	nMOS	Cascode, CC	2x10x1	1:1
(f)	SiGe HBT	Wilson	0.1x10x1	1:1
(g)	SiGe HBT	Cascode	0.1x10x1	1:1
(h)	SiGe HBT	Simple	0.1x2x1 / 0.1x6x2	1:6
(i)	pMOS	CC	1x1x12 / 1x1x6	2:1
(j)	pMOS	CC	0.5x2x2	1:1

CC: common centroid layout

work, CMs will be referred to by the letter in the first column of Table 7.1. The sizing is given in emitter length and width ($\mathrm{L_E} \times \mathrm{W_E} \times \mathrm{N}$) for N SiGe HBTs in parallel, and gate length, width, and the number of fingers ($\mathrm{L_G} \times \mathrm{W_G} \times \mathrm{Fingers}$) for the CMOS. The size of the CMOS transistors was selected based on three criteria: 1) $\mathrm{L_G}$ at least 5 times greater than the minimum in order to improve the output resistance and device matching [125], 2) $\mathrm{W_G}/\mathrm{L_G}$ large enough for adequate current drive, and 3) the total die area of the CMs be comparable to that of SiGe HBT CMs, to ensure a fair comparison. Based on 300 K simulations, the characteristics of the CMOS in the present work are representative of devices with $\mathrm{L_G} \leq 2\mu\mathrm{m}$ (the maximum allowed value for regular$-\mathrm{L_G}$ devices). For SiGe HBTs, the $\mathrm{L_E}$ was selected to cover a small ($2\times$ minimum) and a maximum value. Note that $\mathrm{W_E}$ is fixed in this process technology.

7.3 CM Ratio, Measurement Results, and Discussion

7.3.1 Summary of Expectations

Ideal CMs can mirror an arbitrary level of input current regardless of the output voltage. In realistic CMs, however, the current between the input and the output device can differ when the input and output voltages are different, due to the following effects.

For CMOS, both channel-length modulation (CLM) and drain-induced barrier lowering (DIBL) can result in a different drain current at the same gate-source voltage, when the drain-source voltage ($\mathrm{V_{DS}}$) is different [126, 127]. Based on an early investigation, CLM is dominant for channel length $> 2\mu\mathrm{m}$, while DIBL dominates at smaller lengths [128]. Therefore, all CMOS CMs in the present work, except (e), will be mostly subject to the DIBL effect. At cryogenic temperatures, the DIBL effect can be slightly worse but mostly remains insensitive to temperature [129, 130].

In SiGe HBTs, the Early effect is negligible at room temperature, and even less significant at lower temperatures, due to the Ge grading [35]. However, self-heating is present at high currents and can cause CM errors if the two devices in the core have different collector-emitter voltages (i.e., different power dissipation). Due to their larger thermal resistance, more self-heating is expected at the same current in smaller devices [131]. Self-heating is expected to become weaker, at first, with decreased temperatures, but then increase sharply again below 30-50 K, following the temperature trend of the thermal conductivity of Si [132]. Therefore, CM errors from self-heating can exist at both room temperature and deep cryogenic temperatures, and larger error can be expected for smaller devices. In addition, non-ideal base leakage current, which is frequently observed for cryogenically-operated SiGe HBTs, can degrade current gain (β), which contributes to CM errors. The leakage is believed to originate from the emitter-base oxide interface or depletion region [126]. Usually, smaller devices have more leakage due to a larger perimeter/area ratio [133]. Base leakage can increase CM errors, particularly at low currents, where gain degradation can become significant.

Given this summary of factors that can contribute to CM errors, the next section examines the measured CM error versus input current and output voltage.

7.3.2 Input Current Range

CM ratios versus input current are plotted for various fixed output voltages in Figure 7.1, for 300 K and 19 K. The solid lines denote forward-active bias for SiGe HBTs or saturation bias for CMOS, while the dashed lines represent the nominal CM ratio. The y-axis limit for all CMs is set to 0-200% of the nominal value, to facilitate comparisons. Similar results were obtained from two other chips, but are not shown here for brevity.

At 300 K, all SiGe HBT CMs except (a) exhibit negligible errors across input current. At high current, errors can be seen for (a) and (h) due to different amount of self-heating. Other factors can also contribute to errors, suggested by a larger mismatch for adjacent devices at high current specified by this process technology. For CMOS CMs, since the output voltage is fixed while the input voltage varies with input current, DIBL effects cause large errors in (d, i, j) for most of the range of input current, except when the input and output voltage are equal. For the CM using longer-channel devices (e), the error likely comes more from the CLM than DIBL.

At 19 K, most CMs have increased errors compared to those at 300 K. For SiGe HBT CMs ($a - c$ and $f - h$), large errors can be seen at low injection due to base leakage current. Errors due to self-heating are also visible at high injection for CMs built from small SiGe HBTs (a and h). For CMOS CMs, the errors at 19 K are more dependent on V_{DS} than at 300 K, causing the range of errors to increase. Among all CMs, a few (g, b, c) achieve reasonably small errors across a range of input current (albeit, a smaller range than at 300 K). The implications will be discussed in subsection 7.3.5. In addition, as shown by the solid lines, the error is smaller for (b, c, d, i) when the output device is operating in the forward-active/saturation region for SiGe HBT/CMOS.

An interesting observation is that the SiGe HBT CM (h) constructed with different input and output transistor sizes has extremely large error (>200%). A separate measurement (not shown) reveals that the collector current at cryogenic temperatures does not scale linearly with emitter length, as classically expected and observed at 300 K, producing the

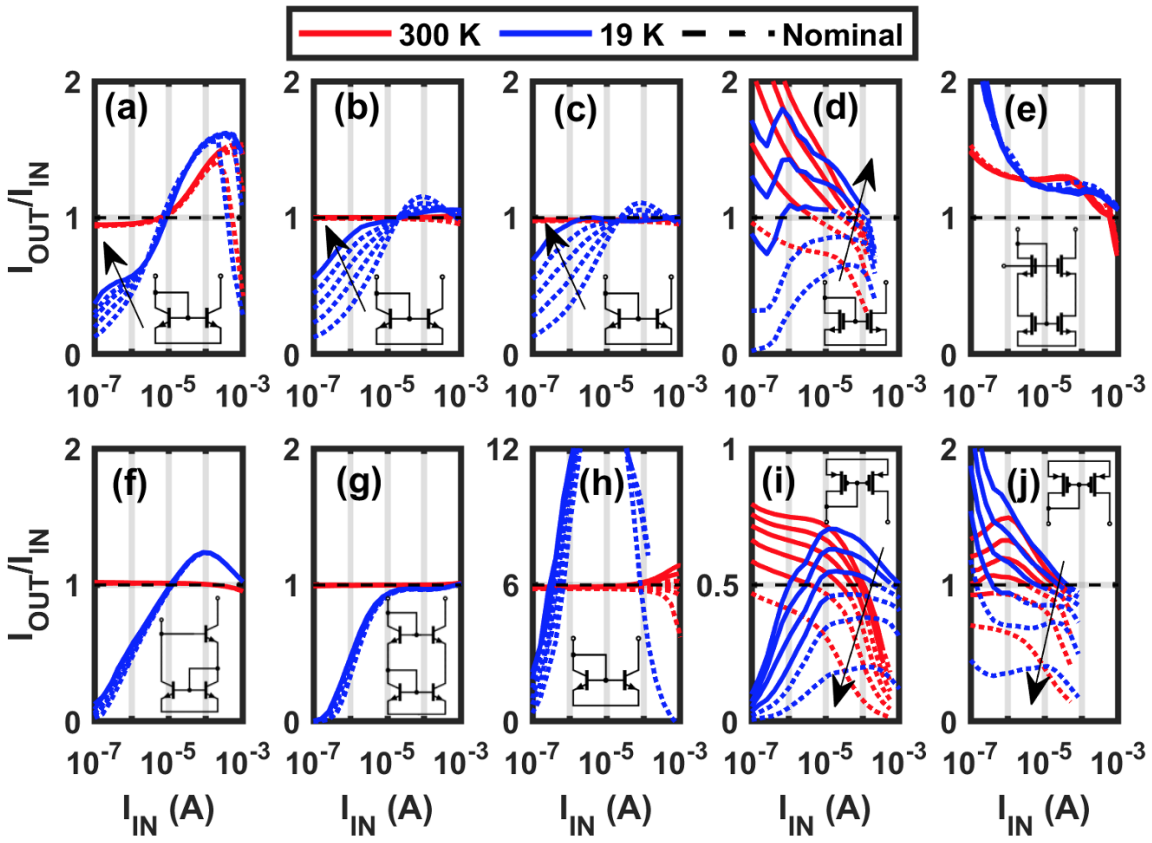

Figure 7.1: Current mirror ratios versus input bias for various fixed output voltages. Figure numbers correspond to the topologies listed in Table 7.1. Arrows point to increasing output voltage. The range of output voltages are: 0.2 to 1 V in 0.2 V step for (a-d) and (h), 0 to 1 V in 0.2 V step for (i-j), and 1.2 to 2.2 V in 0.2 V step for (e-g). Solid lines denote the forward-active region, while dotted lines denote saturation or triode region.

large error in (h). The reason for this non-classical behavior is still under investigation, but clearly CMs with different unit sizes between the input and output devices should be avoided for cryogenic operation.

7.3.3 Output Voltage Range

Another key aspect of any CM is the ability to operate under different output voltages. Although the output voltage is ideally set identical to the input voltage, it can be difficult to achieve in practice. For example, due to the limited voltage headroom and power consumption for cryogenic applications, a lower output voltage is desired. In addition, the input current can change for a variety of reasons, which changes the input voltage, but not

the output voltage, producing CM errors. Even if circuit-feedback is used to equalize the input and the output voltage, more circuitry and chip area is required, an unfavorable trade-off. Overall, it is desirable to have a CM that can maintain small errors under a large range of output voltages. In other words, it should maintain a large output resistance.

The output resistance for all CMs can be examined in Figure 7.2, where the CM ratio versus the output voltage is plotted for two input currents. A larger output resistance means that the curves have a smaller slope. At both 300 K and 19 K, all SiGe HBT CMs ($a - c$, $f - h$) and the CMOS cascode CM (e) have a larger output resistance than the simple CMOS CMs (d, $i - j$). It is well-known that SiGe HBTs have a large output resistance (negligible Early effect) [35]. Though the output resistance can be degraded by the presence of self-heating, the degradation is much less severe than the DIBL effect in CMOS. A cascode topology can be used to increase the output resistance by forcing the lower transistors to maintain the same voltage. Cascode CMs (g, f, e) do achieve high output resistance. However, it should be noted that large output resistance does not necessarily translate to small CM errors. Rather, it simply means the same level of error can be maintained across output voltage. The cascode CMOS CM (e), and the simple CM using small SiGe HBTs (a), both have large output resistance but also sizable errors. The cascode CM with SiGe HBTs (g) has both a large output resistance and small error at both temperatures.

7.3.4 Layout Efficiency

Another aspect of CM design is the layout efficiency. CMOS CMs in the present work are sized similar to the SiGe HBT CMs, and were shown to have significantly worse performance at fixed area. Even if the $\mathrm{W_G}$ and $\mathrm{L_G}$ values in CMOS devices are reversed, based on 300 K simulations, the DIBL and CLM effects at a length of a few $\mu\mathrm{m}$ still introduce more CM error in CMOS CMs than in SiGe HBT CMs, making CMOS CMs less appealing from a performance per chip area perspective.

Within the variety of SiGe HBT CMs, layout efficiency can be compared for (b) and

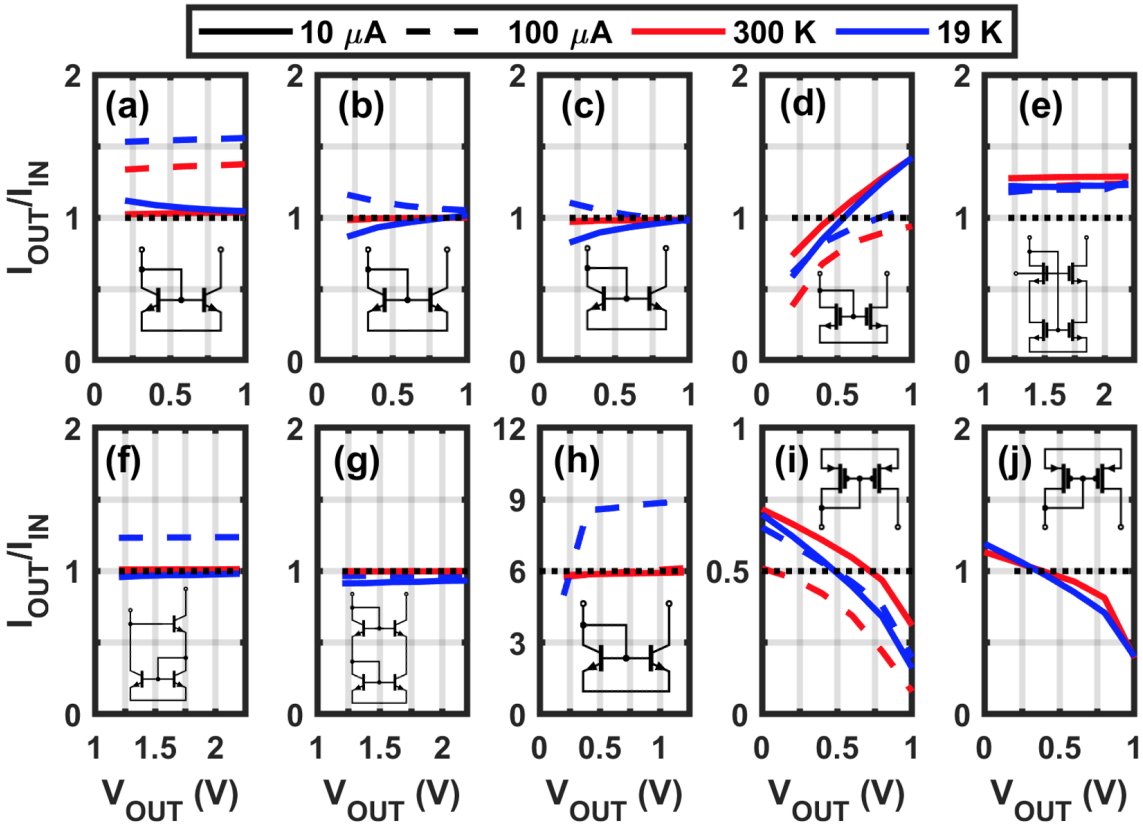

Figure 7.2: Current mirror ratios versus output voltage at a fixed input current of 10 and 100 μA, at 300 K and 19 K.

(c). Both CMs have the same total emitter area of 10 $\mu\mathrm{m}^2$, but (b) uses a single device while (c) uses 8 devices in a common-centroid (CC) layout. At 300 K, both have small errors. At 19 K, the maximum error for b and c (at around 100 μA) is 16.8% versus 11.1%. Although CC layout provides significant improvement, a maximum error of 16.8% without CC layout is already reasonable, given the large errors observed in most CMs at 19 K. In addition, as Section 7.4 will show, the more complicated metal routing associated with the CC layout can potentially cause additional CM errors.

7.3.5 Discussion and Best Practice

As has been shown, SiGe HBT CMs tend to have more robust performance than CMOS CMs at cryogenic temperatures, in terms of both the operable range of input current and operable output voltages. These two advantages allow the same SiGe HBT CM to be used as a versatile topology for a variety of scenarios, without re-design or re-sizing. In comparison, CMOS CMs only achieve low error for a specific bias, requiring a different transistor size when the intended operating condition changes. Moreover, the particular bias at which CMOS CMs achieve low error is subject to change with process variations, further complicating the circuit design. Overall, SiGe HBT CMs offer the performance and flexibility that is highly beneficial for cryogenic circuit design, and should be exploited to designers' advantage, when available.

However, not all SiGe HBT CMs are equally suitable. CMs with small SiGe HBTs are not preferred due to their larger thermal resistance and added base leakage. Cascode SiGe HBT CMs may not be feasible for situations with very limited power consumption. CMs with different unit sizes should also be avoided until the root cause of their large errors is better understood. CC layout is also not necessarily needed in most situations.

Overall, using a simple CM with *large* SiGe HBTs offer the best performance, flexibility, and simplicity. If CMOS CMs have to be used (not recommended, as explained above), the gate length should be maximized. However, since a correspondingly large width is

needed for a particular current drive ($\propto \mathrm{W_G/L_G}$), the length increase should be subject to the overall area constraint. If a large voltage headroom is available, cascode CMOS CMs are strongly preferred over simple CMs, but in that case cascode SiGe HBT CM represents an even better choice in a SiGe BiCMOS technology. For all cases, CM errors tend to deteriorate at cryogenic temperatures and require more care in the design cycle.

7.4 Fundamental Limit of CM Errors

In Section 7.3, different input/output voltages are attributed to be the culprit of observed CM errors. When the voltage is identical, however, there can still be CM errors due to process variations[125] and circuit layout [134]. This section investigates the fundamental limit of CM errors for SiGe HBT CM (b), which has been shown above to offer good overall performance and simplicity of design. The key finding here is that the lowest limit for CM errors also increases (worsens) at decreased temperature due to the increasing effect of process variations and circuit layout. In particular, at lower temperatures, the back-end-of-line (BEOL) layout can have more of an effect than process variations in modifying the observed current. Therefore, a minor difference in layout can potentially contribute to a major CM error. This adds a new dimension of complexity not known to the existing literature.

7.4.1 Test Structures

Two types of structures were measured for this study. The first type is the CM (b) characterized in Section 7.3. The second type is a pair of standalone SiGe HBTs, henceforth referred to as the "device pair". The device pair has identical transistor layout and size (0.1x10 $\mu\mathrm{m}^2$) as in the CM (b), but different BEOL routing, where the CM has more higher-metal traces located above the SiGe HBTs. To minimize the effects of process variations, the two SiGe HBTs are placed at the minimum distance allowed by the design rules. Three chips from the same multi-project-wafer (MPW) run were measured, where three structures (two device

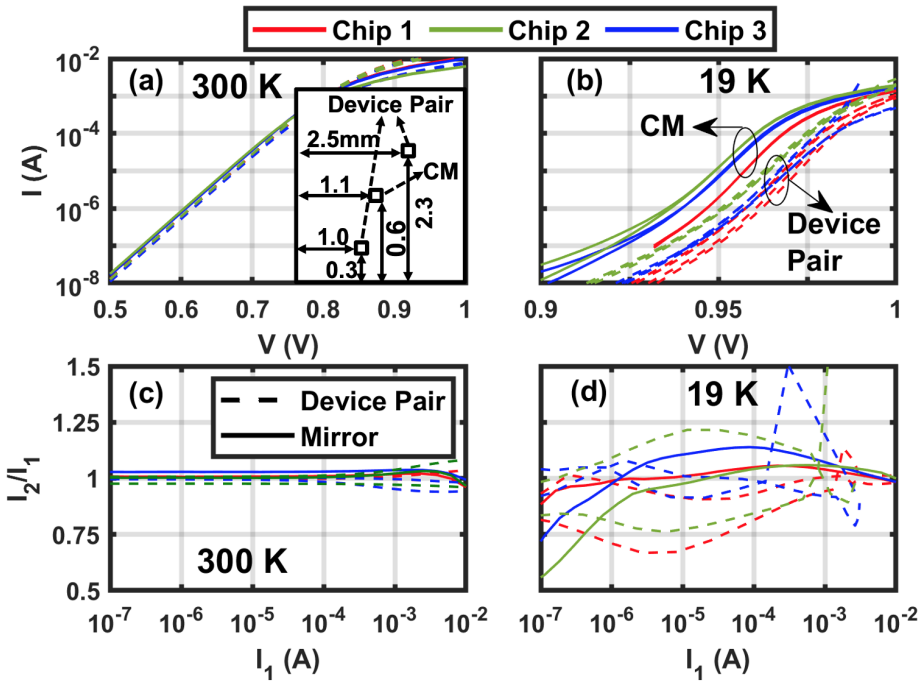

Figure 7.3: (a-b) Current versus voltage for device pairs and CMs at 300 K and 19 K. (c-d) Ratios of currents within each device pair or CM at 300 K and 19 K. Inset of (a): illustration of layout placement of the CM and device pairs in each chip (not drawn to scale).

pairs and one CM) were placed in each chip, as illustrated in the inset of Figure 7.3(a).

7.4.2 Measurement Results

In total, there are 18 nominally identical SiGe HBTs (9 pairs) from 3 chips, making it possible to compare the effects of chip-to-chip and within-chip process variations with the effects of BEOL layout. Towards this goal, current-voltage (I-V) characteristics were measured for all 18 devices. In particular, the I-V characteristics correspond to the input/output currents versus the input voltage for the CM, and collector current versus base-emitter voltage for the device pair. The results are plotted in Figure 7.3(a-b) for 300 K and 19 K. To aid comparisons, Figure 7.3(c-d) shows the ratios of the current between two devices within each pair. The ratio is ideally unity if no device mismatch exists.

As can be seen, all I-V curves nearly overlap at 300 K and the ratio of current within each structure is near unity. This means that the process variations are small across the

three chips, and the difference in BEOL layout (e.g., between the device pair and the CM) has a negligible effect on the current. At 19 K, however, a clear separation of curves can be seen in Figure 7.3(b) between the CMs and the device pairs. In other words, the difference in BEOL results in little difference in current at 300 K, but large differences at 19 K.

To compare the effects of layout difference versus the effects of process variations, the current values for all 18 devices in Figure 7.3(a-b) at a vertical cut-line of 0.63/0.95 V for 300/19 K were gathered, normalized, and plotted in the bar chart shown in Figure 7.4(a-b). The normalization is performed by dividing all values by the smallest value at each temperature, a procedure that simplifies the comparisons between 300 K and 19 K. Figure 7.4(a-b) shows, from left to right for each chip, the two devices from the first device pair, followed by the two devices from the second device pair, and finally the two devices from the CM. At 300 K, device pairs and the CM differ visibly due to the different BEOL layouts. More strikingly, such differences significantly increase at 19 K. Similarly, the chip-to-chip variation is small at 300 K for identical structures on different chips, but increases at 19 K.

To obtain a quantitative picture, the ratios between currents in Figure 7.4(a-b) were calculated. The ratios represent how much the following four effects change the device current: 1) layout differences, 2) chip-to-chip variations, 3) within-chip variations, and 4) residual process variations for two minimum-spaced devices. For example, the ratio of current between the fifth and the first device in chip one corresponds to the effects of BEOL differences on current. As detailed in Table 7.2, many such ratios can be obtained for each effect due to the permutations of 18 devices across 3 chips. The statistics for each effect are plotted as a box plot in Figure 7.4(c-d) for 300 and 19 K. The maximum ratio for each effect is shown above the box.

Immediately, it can be observed that the effects of BEOL layout in changing the device current is the strongest compared to the chip-to-chip and within-chip process variations, at both 300 K and 19 K. In comparison, the minimum-spaced devices have the least difference, although such differences still increase with decreasing temperature, due to the

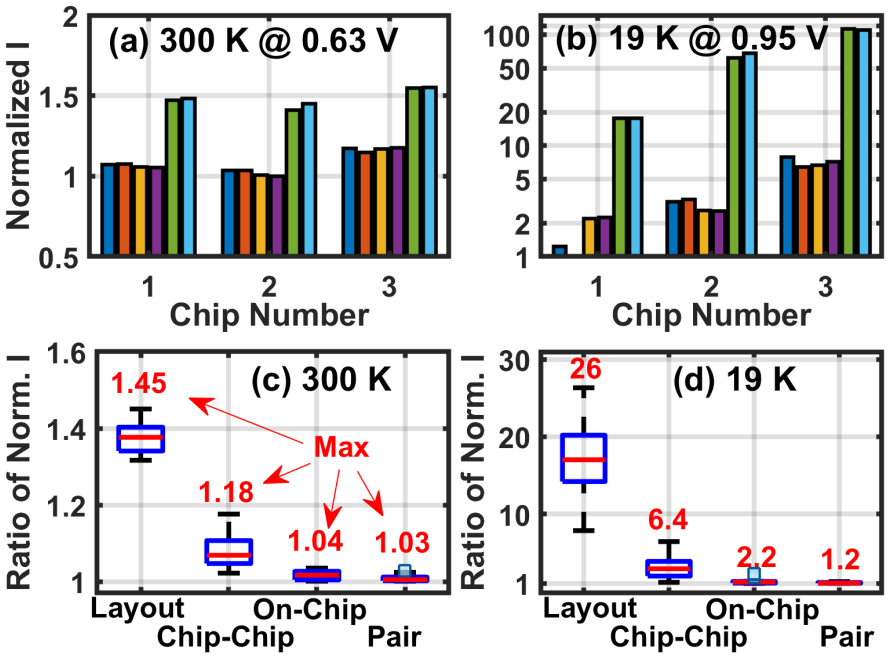

Figure 7.4: (a-b): Current of all devices from Figure 7.3(a) at 0.63 V and from Figure 7.3(b) at 0.95 V. Normalized by the smallest value at each temperature. For each chip, from left to right are: two devices in the device pair 1, in device pair 2, and in the CM. (c-d): Statistics of current ratios obtained by permutations of all possible pairs of values in (a-b). The number of permutations used in the statistics calculation is 24, 18, 21, and 9, respectively, for the four groups from left to right.

increasing effects of process variations at low temperatures, as explained in our separate work [4]. However, the relative difference between the layout effects and chip-to-chip process variations increases drastically from 300 K to 19 K (1.23× versus 4.06×). In other words, given the same layout and process variations, the effects of layout differences dominate in determining the device current, compared to process variations, as temperature decreases.

The physical reason for increasing effects of layout and process variations at lower temperatures is explained in our separate work [4]. Essentially, layout or process variations cause variations in energy bandgap. Since the bandgap shows up in the exponent of the Arrhenius equation, its variation will be exponentially enhanced with decreasing temperature [4]. Clearly, this effect is fundamental and universal for all SiGe HBTs, regardless of technology generation. Although not investigated here, similar layout effects, such as those due to implanted N-wells/P-wells, or shallow-trench isolation (STI), have also been observed for CMOS[135, 136]. Subtle differences in layout, together with the process variations, ultimately limits the lowest achievable errors that CMs at cryogenic temperatures.

7.4.3 Implications for CM Layout

Since BEOL layout has an enhanced effect on device current at cryogenic temperatures, one would think that by making the BEOL layout identical (or symmetrical), its contribution to device mismatch can be reduced. However, for most CMs, it is difficult to achieve a perfectly symmetrical layout. For example, the output device can not be diode-connected, as the input device is, and such a minor difference can potentially contribute to added CM errors. In addition, other circuits that rely on the CM are often placed near the CMs, potentially contributing to the mismatch of the local layout. Based on a room-temperature study [134], layout mismatch up to 40 μm away can still affect CM ratios, a value that will surely worsen with cooling. Therefore, the importance of symmetrical layout cannot be underestimated. Moreover, it is plausible that a portion of the error observed in the

Table 7.2: Details of the Grouping for Figure 7.4

Category	Ratio Details	# Permutations
Layout	(C1,D5) ÷ (C1,D1-4)	4
	(C1,D6) ÷ (C1,D1-4)	4
	(C2,D5) ÷ (C2,D1-4)	4
	(C2,D6) ÷ (C2,D1-4)	4
	(C3,D5) ÷ (C3,D1-4)	4
	(C3,D6) ÷ (C3,D1-4)	4
Total		**24**
Chip-to-Chip	(C2,D1-6) ÷ (C1,D1-6)	6
	(C3,D1-6) ÷ (C1,D1-6)	6
	(C3,D1-6) ÷ (C2,D1-6)	6
Total		**18**
Within-Chip	(C1-3,D2) ÷ (C1-3,D1)	3
	(C1-3,D3) ÷ (C1-3,D1)	3
	(C1-3,D4) ÷ (C1-3,D1)	3
	(C1-3,D3) ÷ (C1-3,D2)	3
	(C1-3,D4) ÷ (C1-3,D2)	3
	(C1-3,D3) ÷ (C1-3,D4)	3
	(C1-3,D6) ÷ (C1-3,D5)	3
Total		**21**
Matched Pair	(C1,D2/4/6) ÷ (C1,D1/3/5)	3
	(C2,D2/4/6) ÷ (C2,D1/3/5)	3
	(C3,D2/4/6) ÷ (C3,D1/3/5)	3
Total		**9**

Notation: (C1,D1-6) denotes device 1 to device 6 on chip 1.
(C2,D1/3/5) denotes device 1, 3, and 5 on chip 2.

previous sections is actually due to the unintentional mismatch in layout.

Overall, an important takeaway is that, in addition to the mismatch of operation conditions, BEOL layout can significantly contribute to CM errors at low temperatures.

7.5 Effect of CM Errors on RF Circuits

Given the increasing CM errors observed at cryogenic temperatures, it is important to investigate its impact on circuit operation. An RF low-noise amplifier (LNA) was chosen as a case study, given its relevance to quantum computing readout applications. Similar effects will occur for mixers and voltage-controlled oscillators, both of which need CMs to bias the core. This investigation is organized as follows. First, a HICUM/L0 [115] compact model for SiGe HBTs operating at 19 K was extracted, using the method outlined in [5]. Next, an LNA was designed for 19 K operation in order to obtain realistic values for passive components and device sizes. Subsequently, variations were introduced to the model parameters of SiGe HBTs to emulate the effects of CM errors on circuit performance. These variations can have different physical origins (e.g., random process variations or unintentional layout differences) but are not distinguished at the circuit level. Finally, the resulting variations of S-parameters and noise temperatures ($\mathrm{T_e}$) of the LNA were obtained from Monte Carlo simulations to evaluate the impact of CM errors on the circuit performance.

The schematic of the designed LNA is shown in Figure 7.5(a). Qualitatively, CM errors shift the bias applied to the core device. Since device parameters (e.g., transconductance, current gain, capacitance, etc.) strongly depend on the bias, a change in bias point results in a change of circuit characteristics. Due to the complex physical origins, CM errors can be modeled as a random variation at the circuit-level, causing random variations in the bias point, and therefore circuit characteristics.

To obtain a quantitative picture, three steps were taken. First, as can be seen in Figure 7.3(b), CM errors can be modeled as a parallel shift of the collector current at a fixed voltage. This is equivalent to the change of the reverse saturation current parameter ($\mathrm{I_S}$)

in the HICUM/L0 model. Therefore, a random Gaussian distribution of I_S was introduced to the bias device to emulate cryogenic CM errors. The remaining devices in the signal path were left unchanged. The standard deviation of the distribution was set to the maximum ratio values (in percentage) for the matched pairs in Figure 7.4(c-d), namely, 2.9% at 300 K and 24% at 19 K. Note that the exact value is less important here, and will likely vary slightly from one technology to another. Key, instead, is the large difference in variability between 300 K and 19 K, which has been universally observed in several BiCMOS technologies [4, 3]. Finally, 250 Monte Carlo simulations were conducted in the Keysight ADS simulator at 19 K and 300 K (at 300 K, using the foundry model at the same bias as at 19 K) to gather the range (minimum and maximum) of the simulated S-parameters and noise temperatures. The results are shown in Figure 7.5(b-d). Although the performance of the LNA is not optimal at 300 K because it has been designed for 19 K operation, the focus here is the range of variations due to CM errors, and not the nominal performance.

As shown in Figure 7.5, the variation of the circuit gain (S_{21}) and the input matching (S_{11}) at 19 K is noticeably larger compared to that at 300 K. A 6-dB difference in gain is present due to the induced CM error. The input matching variation also changes more at 19 K than at 300 K, due to the presence of CM errors. Though the variation in noise temperature is smaller in magnitude at 19 K, well-designed cryogenic SiGe LNAs often have only a few K of noise temperature. Therefore, the additional variation of a few K is significant and can cause some circuits to fail to meet design specifications.

In our previous work [3], DC parameters of SiGe HBTs were found to vary more than RF parameters at cryogenic temperatures. To gain an additional perspective for the severity of CM errors, random variations in junction capacitance were introduced, while keeping I_S fixed. This is effectively the hypothetical case when only RF parameters have variations, and can be contrasted with the previous case with only DC parameter variations (i.e., via the CM errors). The capacitance variation is obtained from a separate characterization (not shown) to be 2.5% and 0.5% for base-emitter capacitance (C_{BE}) and base-collector

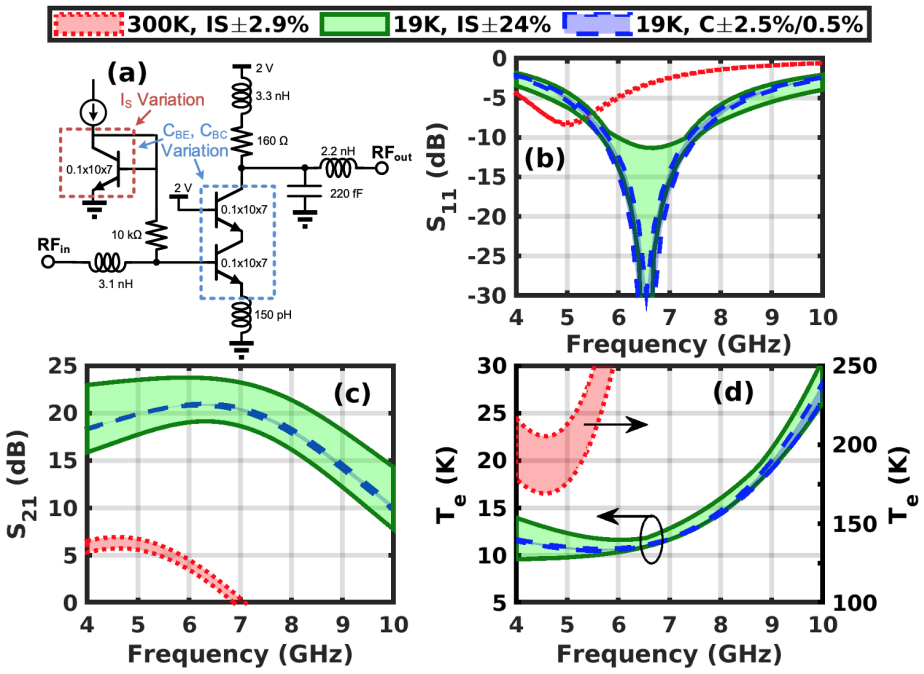

Figure 7.5: (a) Schematic of the LNA under study. (b) Simulated input matching, (c) Simulated gain, and (d) Simulated noise temperatures due to current mirror errors at 300 K and 19 K or due to the device capacitance variations at 19 K.

capacitance (C_{BC}). As shown in Figure 7.5, compared to the I_S variations, the variation in gain is much smaller with capacitance variations. This is because the gain is mostly limited by the DC transconductance, which varies exponentially with the collector current. The variation of input matching and noise temperature due to capacitance variation is mostly negligible compared to that due to I_S variation.

Overall, it is found that CM errors lead to large uncertainty in circuit S-parameters and noise performance due to induced changes in bias point. With increasing CM errors at cryogenic temperatures, this uncertainty increases, which makes the circuit more prone to falling outside design specifications, thus potentially decreasing yield. In addition, the variation of DC parameters (and the resulting CM errors) can induce much larger variations in performance than RF parameters.

7.6 Summary

The present work demonstrates that the errors of most CM topologies increase at cryogenic temperatures. The reasons for increasing CM errors are: 1) the mismatch in bias conditions and 2) the increasing effects of circuit layout and process variations. The effect of BEOL layout on device current increases drastically at cryogenic temperatures, making it a top concern for accurate cryogenic CM design. Depending on the usage scenario, simple CMs using large SiGe HBTs offer small errors, a large operational range, and a simple layout, while cascode CMs using SiGe HBTs offer both low errors and high output resistance. Common-centroid layout is found to offer marginal improvement. SiGe HBT CMs with different unit sizes between input and output devices should be avoided. Due to the increasing CM errors at lower temperatures, the variation of circuit performance also increases due to bias point changes, potentially leading to more post-fabrication circuits falling outside the intended specifications.

CHAPTER 8
CONCLUSIONS

8.1 Summary of Contributions

This thesis investigates several aspects of deep cryogenic operations of SiGe HBTs relevant for the emerging applications in quantum science. The operation of SiGe HBTs at deep cryogenic temperatures are characterized, studied, and modeled, and simple circuits are used to verify the findings. Overall, the findings clarify several important unresolved questions and clear the roadblocks of using SiGe HBTs for quantum science. The knowledge from this thesis, combined with the existing knowledge from the CMOS community, paves the foundation for using BiCMOS technology for future high-performance circuits demanded by quantum science applications.

The contributions of this research include:

1. Demonstration successful transistor operations of SiGe HBTs at a record-low temperature of 70 mK and a record high magnetic field of 14 T

2. A method to distinguish the newly discovered direct tunneling mechanism from quasi-ballistic transport

3. Prediction of the effect of future technology scaling on device performance

4. Discovery of the increasing variability for DC and RF parameters in SiGe HBTs at cryogenic temperatures

5. Physical origin of the increasing variability of SiGe HBTs at cryogenic temperatures and possible mitigation methods

6. A compact model constructed for cryogenically-operated SiGe HBTs

7. Discovery of increasing current mirror errors at lower temperatures due to device variability

8. Demonstration of increasing variability of circuit performance due to device variability

8.2 Future Work

The field of using SiGe HBTs and BiCMOS technology for quantum science is exciting and open. This thesis paves a foundation for many interesting topics for further research and demonstration, such as

1. Device structure and profile optimization that will exploit the presence of direct tunneling for possible performance improvement at cryogenic temperatures

2. Device profile engineering that reduces the enhancement of variability due to lowering temperature

3. Demonstration of mixers, LNAs, VCOs, and other circuit blocks relevant for quantum science application

4. Extension of the compact model to have a more accurate temperature dependence, transistor-size dependence, and large-signal modeling

5. Cryogenic compact model of passives (inductor, capacitor, resistors), along with their statistics modeling

6. Performance comparison of circuits built from SiGe HBTs and those built from advanced-node CMOS at cryogenic temperatures for a cost-performance analysis

Appendices

APPENDIX A

VARIABILITY OF $\mathrm{EXP}(-E_A/KT)$

The Arrhenius equation, $e^{-E_a/kT}$ is often encountered in device equations, where E_a is the activation energy, k is Boltzmann constant ($= 8.617 \times 10^{-5}$ eV/K), and T is the absolute temperature. We show here that given the same variability of E_a, the variability of $e^{-E_a/kT}$ strongly increases at low temperatures.

If E_a follows a normal distribution, $e^{-E_a/kT}$ will have a log-normal distribution. The relative standard deviation of $e^{-E_a/kT}$ can be obtained as [137]

$$RSD(e^{-E_a/kT}) = \sqrt{\exp\left[\left(\frac{\sigma_{E_a}}{kT}\right)^2\right] - 1} \propto e^{\sigma_{E_a}/kT}$$

where σ_{E_a} is the standard deviation of E_a. As can be seen, σ_{E_a} is multiplied by the factor of $1/kT$ in the exponent, which becomes larger with decreasing temperature. Numerically, $1/(8.617\times10^{-5}\times T) = 11605/T$. At T=100 K, this equals 116, an enormous multiplicative factor. In other words, the same variation of E_a leads to a larger variation of $e^{-E_a/kT}$ at low temperatures.

REFERENCES

[1] H. Ying, B. R. Wier, J. Dark, N. E. Lourenco, L. Ge, A. P. Omprakash, M. Mourigal, D. Davidovic, and J. D. Cressler, "Operation of SiGe HBTs down to 70 mK," *IEEE EDL*, vol. 38, no. 1, pp. 12–15, 2017.

[2] H. Ying, J. Dark, A. P. Omprakash, B. R. Wier, L. Ge, U. Raghunathan, N. E. Lourenco, Z. E. Fleetwood, M. Mourigal, D. Davidovic, and J. D. Cressler, "Collector transport in SiGe HBTs operating at cryogenic temperatures," *IEEE Trans. Electron Devices*, vol. 65, no. 9, pp. 3697–3703, Sep. 2018.

[3] H. Ying, J. W. Teng, G. N. Tzintzarov, A. P. Omprakash, S. G. Rao, U. Raghunathan, A. Ildefonso, M. S. Fernandez, and J. D. Cressler, "DC and RF variability of SiGe HBTs operating down to deep cryogenic temperatures," in *2019 IEEE BiCMOS and Compound Semicond. Integr. Circuits and Technol. Symp. (BCICTS)*, Nov. 2019.

[4] H. Ying, J. W. Teng, U. S. Raghunathan, J. P. Moody, and J. D. Cressler, "Variability of p-n junctions and sige hbts at cryogenic temperatures," *IEEE Trans. Electron Devices*, accepted, 1/20/2021.

[5] H. Ying, S. G. Rao, J. W. Teng, M. Frounchi, M. Muller, X. Jin, M. Schroter, and J. D. Cressler, "Compact modeling of SiGe HBTs for design of cryogenic control and readout circuits for quantum computing," in *2020 IEEE BiCMOS and Compound Semicond. Integr. Circuits and Technol. Symp. (BCICTS)*, Nov. 2020.

[6] H. Ying, J. W. Teng, and J. D. Cressler, "Operation of current mirror in SiGe BiCMOS technology at cryogenic temperatures," *IEEE Trans. Electron Devices*, accepted, 2/15/21.

[7] V. Vedral *et al.*, *Introduction to quantum information science*. Oxford University Press on Demand, 2006.

[8] M. A. Nielsen and I. Chuang, *Quantum computation and quantum information*. American Association of Physics Teachers, 2002.

[9] A. Montanaro, "Quantum algorithms: an overview," *npj Quantum Information*, vol. 2, p. 15 023, 2016.

[10] R. De Wolf, "The potential impact of quantum computers on society," *Ethics and Information Technology*, vol. 19, no. 4, pp. 271–276, 2017.

[11] S. J. Devitt, W. J. Munro, and K. Nemoto, "Quantum error correction for beginners," *Reports on Progress in Physics*, vol. 76, no. 7, p. 076 001, 2013.

[12] G. G. Guerreschi and A. Y. Matsuura, "QAOA for max-cut requires hundreds of qubits for quantum speed-up," *Scientific reports*, vol. 9, no. 1, p. 6903, 2019.

[13] B. Patra, R. M. Incandela, J. P. G. van Dijk, H. A. R. Homulle, L. Song, M. Shahmohammadi, R. B. Staszewski, A. Vladimirescu, M. Babaie, F. Sebastiano, and E. Charbon, "Cryo-CMOS circuits and systems for quantum computing applications," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 309–321, Jan. 2018.

[14] E. A. Dauler, M. E. Grein, A. J. Kerman, F. Marsili, S. Miki, S. W. Nam, M. D. Shaw, H. Terai, V. B. Verma, and T. Yamashita, "Review of superconducting nanowire single-photon detector system design options and demonstrated performance," *Optical Engineering*, vol. 53, no. 8, p. 081 907, 2014.

[15] C. M. Natarajan, M. G. Tanner, and R. H. Hadfield, "Superconducting nanowire single-photon detectors: physics and applications," *Superconductor science and technology*, vol. 25, no. 6, p. 063 001, 2012.

[16] E. G. Rieffel and W. H. Polak, *Quantum computing: A gentle introduction*. MIT Press, 2011.

[17] J. C. Bardin, D. Sank, O. Naaman, and E. Jeffrey, "Quantum computing: an introduction for microwave engineers," *IEEE Microwave Magazine*, vol. 21, no. 8, pp. 24–44, 2020.

[18] F. Sebastiano, H. Homulle, B. Patra, R. Incandela, J. van Dijk, L. Song, M. Babaie, A. Vladimirescu, and E. Charbon, "Cryo-CMOS electronic control for scalable quantum computing," in *Proceedings of the 54th Annual Design Automation Conference 2017*, 2017, pp. 1–6.

[19] D. Rist, S. Fallek, B. Donovan, and T. A. Ohki, "Microwave techniques for quantum computers: state-of-the-art control systems for quantum processors," *IEEE Microwave Magazine*, vol. 21, no. 8, pp. 60–71, 2020.

[20] J. P. G. Van Dijk, B. Patra, S. Subramanian, X. Xue, N. Samkharadze, A. Corna, C. Jeon, F. Sheikh, E. Juarez-Hernandez, B. P. Esparza, H. Rampurawala, B. R. Carlton, S. Ravikumar, C. Nieva, S. Kim, H. J. Lee, A. Sammak, G. Scappucci, M. Veldhorst, L. M. K. Vandersypen, E. Charbon, S. Pellerano, M. Babaie, and F. Sebastiano, "A scalable Cryo-CMOS controller for the wideband frequency-multiplexed control of spin qubits and transmons," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 11, pp. 2930–2946, 2020.

[21] S. K. Moore, *Google builds circuit to solve one of quantum computings biggest problems*, https://spectrum.ieee.org/tech-talk/semiconductors/design/google-team-builds-circuit-to-solve-one-of-quantum-computings-biggest-problems, Accessed: 2021-1-2.

[22] E. Schroeder, P. Mauskopf, G. Pilyavsky, A. Sinclair, N. Smith, S. Bryan, H. Mani, D. Morozov, K. Berggren, D. Zhu, *et al.*, "On the measurement of intensity correlations from laboratory and astronomical sources with SPADs and SNSPDs," in *Optical and Infrared Interferometry and Imaging V*, International Society for Optics and Photonics, vol. 9907, 2016, 99070P.

[23] B. Korzh, Q.-Y. Zhao, J. P. Allmaras, S. Frasca, T. M. Autry, E. A. Bersin, A. D. Beyer, R. M. Briggs, B. Bumble, M. Colangelo, G. M. Crouch, A. E. Dane, T. Gerrits, A. E. Lita, F. Marsili, G. Moody, C. Peña, E. Ramirez, J. D. Rezac, N. Sinclair, M. J. Stevens, A. E. Velasco, V. B. Verma, E. E. Wollman, S. Xie, D. Zhu, P. D. Hale, M. Spiropulu, K. L. Silverman, R. P. Mirin, S. W. Nam, A. G. Kozorezov, M. D. Shaw, and K. K. Berggren, "Demonstration of sub-3 ps temporal resolution with a superconducting nanowire single-photon detector," *Nature Photonics*, vol. 14, no. 4, pp. 250–255, 2020.

[24] L. He, L. Li, X. Niu, H. Xia, M. Xie, X. Wu, L. Zhang, L. Kang, and Z. Wang, "A low-power, inductorless wideband cryogenic amplifier for superconducting nanowire single photon detector," *IEEE Transactions on Applied Superconductivity*, vol. 29, no. 6, pp. 1–6, 2019.

[25] R. J. Collins, R. Hadfield, V. Fernandez, S. W. Nam, and G. S. Buller, "Low timing jitter detector for gigahertz quantum key distribution," *Electronics Letters*, vol. 43, no. 3, pp. 180–182, 2007.

[26] N. C. Harris, D. Grassani, A. Simbula, M. Pant, M. Galli, T. Baehr-Jones, M. Hochberg, D. Englund, D. Bajoni, and C. Galland, "Integrated source of spectrally filtered correlated photons for large-scale quantum photonic systems," *Physical Review X*, vol. 4, no. 4, p. 041 047, 2014.

[27] A. McCarthy, N. J. Krichel, N. R. Gemmell, X. Ren, M. G. Tanner, S. N. Dorenbos, V. Zwiller, R. H. Hadfield, and G. S. Buller, "Kilometer-range, high resolution depth imaging via 1560 nm wavelength single-photon detection," *Optics express*, vol. 21, no. 7, pp. 8904–8915, 2013.

[28] J. Wu, L. You, S. Chen, H. Li, Y. He, C. Lv, Z. Wang, and X. Xie, "Improving the timing jitter of a superconducting nanowire single-photon detection system," *Appl. Opt.*, vol. 56, no. 8, pp. 2195–2200, 2017.

[29] Q. Zhao, L. Zhang, T. Jia, L. Kang, W. Xu, J. Chen, and P. Wu, "Intrinsic timing jitter of superconducting nanowire single-photon detectors," *Applied Physics B*, vol. 104, no. 3, pp. 673–678, 2011.

[30] N. Calandri, Q.-Y. Zhao, D. Zhu, A. Dane, and K. K. Berggren, "Superconducting nanowire detector jitter limited by detector geometry," *Applied Physics Letters*, vol. 109, no. 15, p. 152 601, 2016. eprint: https://doi.org/10.1063/1.4963158.

[31] D. F. Santavicca, B. Noble, C. Kilgore, G. A. Wurtz, M. Colangelo, D. Zhu, and K. K. Berggren, "Jitter characterization of a dual-readout SNSPD," *IEEE Transactions on Applied Superconductivity*, vol. 29, no. 5, pp. 1–4, 2019.

[32] Y. Hochberg, I. Charaev, S.-W. Nam, V. Verma, M. Colangelo, and K. K. Berggren, "Detecting sub-GeV dark matter with superconducting nanowires," *Physical review letters*, vol. 123, no. 15, p. 151 802, 2019.

[33] OST mission concept study team, "The origins space telescope (OST) mission concept study interim report," *arXiv preprint arXiv:1809.09702*, 2018.

[34] E. Preisler, "A commercial foundry perspective of SiGe BiCMOS process technologies," in *2020 IEEE BiCMOS and Compound Semicond. Integr. Circuits and Technol. Symp. (BCICTS)*, Nov. 2020.

[35] J. D. Cressler and G. Niu, *Silicon-Germanium Heterojunction Bipolar Transistors*. Boston, MA: Artech House, 2002.

[36] J. D. Cressler, "On the potential of SiGe HBTs for extreme environment electronics," *Proceedings of the IEEE*, vol. 93, no. 9, pp. 1559–1582, 2005.

[37] J. Yuan, J. D. Cressler, R. Krithivasan, T. Thrivikraman, M. H. Khater, D. C. Ahlgren, A. J. Joseph, and J. S. Rieh, "On the performance limits of cryogenically operated SiGe HBTs and its relation to scaling for terahertz speeds," *IEEE Transactions on Electron Devices*, vol. 56, no. 5, pp. 1007–1019, 2009.

[38] B. Banerjee, S. Venkataraman, C.-H. Lee, and J. Laskar, "Broadband noise modeling of SiGe HBT under cryogenic temperatures," in *2007 IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, IEEE, 2007, pp. 765–768.

[39] J. D. Cressler and A. Mantooth, *Extreme Environment Electronics*. CRC Press, 2012.

[40] P. S. Chakraborty, A. S. Cardoso, B. R. Wier, A. P. Omprakash, J. D. Cressler, M. Kaynak, and B. Tillack, "A 0.8 THz SiGe HBT operating at 4.3 K," *IEEE Electron Device Letters*, vol. 35, no. 2, pp. 151–153, 2014.

[41] J. C. Bardin, E. Jeffrey, E. Lucero, T. Huang, O. Naaman, R. Barends, T. White, M. Giustina, D. Sank, P. Roushan, K. Arya, B. Chiaro, J. Kelly, J. Chen, B. Burkett, Y. Chen, A. Dunsworth, A. Fowler, B. Foxen, C. Gidney, R. Graff, P. Klimov, J. Mutus, M. McEwen, A. Megrant, M. Neeley, C. Neill, C. Quintana, A. Vainsencher, H. Neven, and J. Martinis, "29.1 a 28nm bulk-CMOS 4-to-8GHz 2mW cryogenic pulse modulator for scalable quantum computing," in *2019 IEEE ISSCC*, 2019, pp. 456–458.

[42] S. Finn, J. Yuan, R. Krithivasan, L. Najafizadeh, P. Cheng, and J. D. Cressler, "A 10 Mbps SiGe BiCMOS transceiver for operation down to cryogenic temperatures," in *Proceedings of 2007 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, 2007, pp. 115–118.

[43] Y. Yao, F. Dai, R. C. Jaeger, and J. D. Cressler, "A 12-bit cryogenic and radiation-tolerant digital-to-analog converter for aerospace extreme environment applications," *IEEE Transactions on Industrial Electronics*, vol. 55, no. 7, pp. 2810–2819, 2008.

[44] J. C. Bardin, P. Ravindran, S. W. Chang, R. Kumar, J. A. Stern, M. D. Shaw, D. S. Russell, and W. H. Farr, "A high-speed cryogenic SiGe channel combiner IC for large photon-starved SNSPD arrays," in *Proceedings of 2013 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, 2013, pp. 215–218.

[45] C. Ulaganathan, N. Nambiar, B. Prothro, R. Greenwell, S. Chen, B. J. Blalock, C. L. Britton, M. N. Ericson, H. Hoang, R. Broughton, K. Cornett, G. Fu, H. A. Mantooth, J. D. Cressler, and R. W. Berger, "A SiGe BiCMOS instrumentation channel for extreme environment applications," in *Proceedings of the 51st Midwest Symposium on Circuits and Systems*, 2008, pp. 217–220.

[46] A. J. Joseph, J. D. Cressler, and D. M. Richey, "Operation of SiGe heterojunction bipolar transistors in the liquid-helium temperature regime," *IEEE Electron Device Letters*, vol. 16, no. 6, pp. 268–270, 1995.

[47] L. Najafizadeh, J. S. Adams, S. D. Phillips, K. A. Moen, J. D. Cressler, S. Aslam, T. R. Stevenson, and R. M. Meloy, "Sub-1-K operation of SiGe transistors and circuits," *IEEE Electron Device Letters*, vol. 30, no. 5, pp. 508–510, 2009.

[48] J. D. Cressler, *The Silicon Heterostructure Handbook: Materials, Fabrication, Devices, Circuits, and Applications of SiGe and Si Strained-Layer Epitaxy*. New York, NY: CRC Press, 2005.

[49] J. D. Cressler, "SiGe-base bipolar transistors for cryogenic BiCMOS applications," in *Proceedings of the 22nd European Solid State Device Research Conference (ESSDERC)*, 1992, pp. 841–848.

[50] W.-T. Wong, M. Hosseini, H. Rucker, and J. C. Bardin, "A 1 mW cryogenic LNA exploiting optimized SiGe HBTs to achieve an average noise temperature of 3.2 K from 4–8 GHz," in *2020 IEEE/MTT-S Int. Microw. Symp. (IMS)*, Aug. 2020.

[51] M Babaie, E. Charbon, A. Vladimirescu, F. Sebastiano, *et al.*, "Characterization and modeling of mismatch in Cryo-CMOS," *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 263–273, 2020.

[52] A. Beckers, F. Jazaeri, H. Bohuslavskyi, L. Hutin, S. De Franceschi, and C. Enz, "Design-oriented modeling of 28 nm FDSOI CMOS technology down to 4.2 K for quantum computing," in *2018 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, IEEE, 2018, pp. 1–4.

[53] H Bohuslavskyi, A. Jansen, S Barraud, V Barral, M Cassé, L Le Guevel, X Jehl, L Hutin, B Bertrand, G Billiot, *et al.*, "Cryogenic subthreshold swing saturation in fd-soi mosfets described with band broadening," *IEEE Electron Device Letters*, vol. 40, no. 5, pp. 784–787, 2019.

[54] D. Davidović, H. Ying, J. Dark, B. Wier, L. Ge, N. Lourenco, A. Omprakash, M. Mourigal, and J. Cressler, "Tunneling, current gain, and transconductance in silicon-germanium heterojunction bipolar transistors operating at millikelvin temperatures," *Phys. Rev. Appl.*, vol. 8, no. 2, Aug. 2017.

[55] H. Ying, "Collector current transport mechanisms in SiGe HBTs operating at cryogenic temperatures," Master's thesis, Georgia Institute of Technology, 2019.

[56] C. K. Andersen, J. Kerckhoff, K. W. Lehnert, B. J. Chapman, and K. Mølmer, "Closing a quantum feedback loop inside a cryostat: autonomous state preparation and long-time memory of a superconducting qubit," *Phys. Rev. A*, vol. 93, p. 012 346, 1 2016.

[57] A. A. Prager, H. C. George, A. O. Orlov, and G. L. Snider, "Experimental demonstration of hybrid CMOS-single electron transistor circuits," *Journal of Vacuum Science & Technology B*, vol. 29, no. 4, p. 041 004, 2011.

[58] K. Das and T. Lehmann, "A cryogenic single electron transistor readout circuit: practical issues and measurement considerations," in *2012 IEEE International Symposium on Circuits and Systems*, 2012, pp. 1359–1362.

[59] A. Robinson and V. Talyanskii, "Cryogenic amplifier for 1 MHz with a high input impedance using a commercial pseudomorphic high electron mobility transistor," *Review of Scientific Instruments*, vol. 75, pp. 3169–3176, 2004.

[60] I. Vink, T Nooitgedagt, R. Schouten, L. Vandersypen, and W. Wegscheider, "Cryogenic amplifier for fast real-time detection of single-electron tunneling," *Applied Physics Letters*, vol. 91, no. 12, p. 123 512, 2007.

[61] A. T. Lee, "Broadband cryogenic preamplifiers incorporating gaas mesfets for use with low-temperature particle detectors," *Review of Scientific Instruments*, vol. 60, no. 10, pp. 3315–3322, 1989.

[62] M. J. Curry, T. D. England, N. C. Bishop, G. Ten-Eyck, J. R. Wendt, T. Pluym, M. P. Lilly, S. M. Carr, and M. S. Carroll, "Cryogenic preamplification of a single-electron-transistor using a silicon-germanium heterojunction-bipolar-transistor," *Appl. Phys. Lett.*, vol. 106, no. 20, p. 203 505, May 2015.

[63] B. Heinemann, R. Barth, D. Bolze, J. Drews, G. G. Fischer, A. Fox, O. Fursenko, T. Grabolla, U. Haak, D. Knoll, R. Kurps, M. Lisker, S. Marschmeyer, H. Rücker, D. Schmidt, J. Schmidt, M. A. Schubert, B. Tillack, C. Wipf, D. Wolansky, and Y. Yamamoto, "SiGe HBT technology with f_T/f_{MAX} of 300GHz/500GHz and 2.0 ps CML gate delay," in *Technical Digest of IEEE International Electron Devices Meeting (IEDM)*, 2010, pp. 30.5.1–30.5.4.

[64] J. J. Pekarik, J. Adkisson, P. Gray, Q. Liu, R. Camillo-Castillo, M. Khater, V. Jain, B. Zetterlund, A. DiVergilio, X. Tian, A. Vallett, J. Ellis-Monaghan, B. J. Gross, P. Cheng, V. Kaushal, Z. He, J. Lukaitis, K. Newton, M. Kerbaugh, N. Cahoon, L. Vera, Y. Zhao, J. R. Long, A. Valdes-Garcia, S. Reynolds, W. Lee, B. Sadhu, and D. Harame, "A 90nm SiGe BiCMOS technology for mm-Wave and high-performance analog applications," in *2014 IEEE BCTM*, 2014, pp. 92–95.

[65] D. M. Richey, A. J. Joseph, J. D. Cressler, and R. C. Jaeger, "Evidence for non-equilibrium base transport in Si and SiGe bipolar transistors at cryogenic temperatures," *Solid-State Electronics*, vol. 39, no. 6, pp. 785–789, 1996.

[66] J. C. Bardin, "Silicon-germanium heterojunction bipolar transistors for extremely low-noise applications," PhD thesis, California Institute of Technology, 2009.

[67] Z. Xu, G. Niu, L. Luo, and J. Cressler, "A physics-based trap-assisted tunneling current model for cryogenic temperature compact modeling of sige hbts," *ECS Transactions*, vol. 33, no. 6, pp. 301–310, 2010.

[68] D. C. Ralph, C. T. Black, and M. Tinkham, "Spectroscopic measurements of discrete electronic states in single metal particles," *Physical Review Letters*, vol. 74, no. 16, pp. 3241–3244, 1995.

[69] E. Bonet, M. M. Deshmukh, and D. C. Ralph, "Solving rate equations for electron tunneling via discrete quantum states," *Physical Review B*, vol. 65, no. 4, 2002.

[70] S. Gustavsson, I. Shorubalko, R. Leturcq, S. Schn, and K. Ensslin, "Measuring current by counting electrons in a nanowire quantum dot," *Applied Physics Letters*, vol. 92, no. 15, p. 152 101, 2008. eprint: https://doi.org/10.1063/1.2892679.

[71] T. Mller, B. Kng, S. Hellmller, P. Studerus, K. Ensslin, T. Ihn, M. Reinwald, and W. Wegscheider, "An in situ tunable radio-frequency quantum point contact," *Applied Physics Letters*, vol. 97, no. 20, p. 202 104, 2010. eprint: https://doi.org/10.1063/1.3517483.

[72] L.-J. Wang, G. Cao, T. Tu, H.-O. Li, C. Zhou, X.-J. Hao, Z. Su, G.-C. Guo, H.-W. Jiang, and G.-P. Guo, "A graphene quantum dot with a single electron transistor as an integrated charge sensor," *Applied Physics Letters*, vol. 97, no. 26, p. 262 113, 2010. eprint: https://doi.org/10.1063/1.3533021.

[73] SHI Cryogenics Group, *4K cryocoolers*.

[74] A. J. Matthews, M Patton, T Marsh, and H van der Vliet, "A compact cryogen-free platform operating at 1 K or 50 mK," *Journal of Physics: Conference Series*, vol. 969, no. 1, p. 012 091, 2018.

[75] ICEoxford, *Cryogen free dilution refrigerator system*.

[76] H. Rücker, J. Korn, and J. Schmidt, "Operation of SiGe HBTs at cryogenic temperatures," in *2017 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, 2017, pp. 17–20.

[77] D. C. Ahlgren, G. Freeman, S. Subbanna, R. Groves, D. Greenberg, J. Malinowski, D. Nguyen-Ngoc, S. J. Jeng, K. Stein, K. Schonenberg, D. Kiesling, B. Martin, S. Wu, D. Harame, and B. Meyerson, "A SiGe HBT BiCMOS technology for mixed signal rf applications," in *Proceedings of the 1997 Bipolar/BiCMOS Circuits and Technology Meeting*, 1997, pp. 195–197.

[78] R. A. Camillo-Castillo, Q. Z. Liu, J. W. Adkisson, M. H. Khater, P. B. Gray, V. Jain, R. K. Leidy, J. J. Pekarik, J. P. Gambino, B. Zetterlund, C. Willets, C. Parrish, S. U. Engelmann, A. M. Pyzyna, P. Cheng, and D. L. Harame, "Sige hbts in 90nm bicmos technology demonstrating 300ghz/420ghz ft/fmax through reduced rb and ccb parasitics," in *2013 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, 2013, pp. 227–230.

[79] B. Jagannathan, M. Khater, F. Pagette, J. S. Rieh, D. Angell, H. Chen, J. Florkey, F. Golan, D. R. Greenberg, R. Groves, S. J. Jeng, J. Johnson, E. Mengistu, K. T. Schonenberg, C. M. Schnabel, P. Smith, A. Stricker, D. Ahlgren, G. Freeman, K. Stein, and S. Subbanna, "Self-aligned SiGe NPN transistors with 285 GHz f_{MAX} and 207 GHz f_T in a manufacturable technology," *IEEE Electron Device Letters*, vol. 23, no. 5, pp. 258–260, 2002.

[80] J. J. Pekarik, J. Adkisson, P. Gray, Q. Liu, R. Camillo-Castillo, M. Khater, V. Jain, B. Zetterlund, A. DiVergilio, X. Tian, A. Vallett, J. Ellis-Monaghan, B. J. Gross, P. Cheng, V. Kaushal, Z. He, J. Lukaitis, K. Newton, M. Kerbaugh, N. Cahoon, L. Vera, Y. Zhao, J. R. Long, A. Valdes-Garcia, S. Reynolds, W. Lee, B. Sadhu, and D. Harame, "A 90nm SiGe BiCMOS technology for mm-wave and high-performance analog applications," in *2014 IEEE Bipolar/BiCMOS Circuits and Technol. Meeting (BCTM)*, Sep. 2014.

[81] S. Tanaka and M. S. Lundstrom, "A flux-based study of carrier transport in thin-base diodes and transistors," *IEEE Transactions on Electron Devices*, vol. 42, no. 10, pp. 1806–1815, 1995.

[82] M. A. Stettler and M. S. Lundstrom, "A microscopic study of transport in thin base silicon bipolar transistors," *IEEE Transactions on Electron Devices*, vol. 41, no. 6, pp. 1027–1033, 1994.

[83] A. A. Grinberg and S. Luryi, "Diffusion in a short base," *Solid-State Electronics*, vol. 35, no. 9, pp. 1299 –1309, 1992.

[84] E Merzbacher, *Quantum mechanics, 1998*.

[85] M. Schroter, G. Wedel, B. Heinemann, C. Jungemann, J. Krause, P. Chevalier, and A. Chantre, "Physical and electrical performance limits of high-speed SiGeC HBTs part I: vertical scaling," *IEEE Transactions on Electron Devices*, vol. 58, no. 11, pp. 3687–3696, 2011.

[86] D. Klaassen, J. Slotboom, and H. de Graaff, "Unified apparent bandgap narrowing in n- and p-type silicon," *Solid-State Electron.*, vol. 35, no. 2, pp. 125–129, Feb. 1992.

[87] D. Klaassen, "A unified mobility model for device simulationi. model equations and concentration dependence," *Solid-State Electronics*, vol. 35, no. 7, pp. 953–959, 1992.

[88] *Sentaurus device user guide*, version Version N-2017.09, Synopsys Inc., September, 2017, 1546 pp.

[89] M. Veldhorst, H. G. J. Eenink, C. H. Yang, and A. S. Dzurak, "Silicon CMOS architecture for a spin-based quantum computer," *Nature Communications*, vol. 8, no. 1, p. 1766, 2017.

[90] E. Charbon, F. Sebastiano, M. Babaie, A. Vladimirescu, M. Shahmohammadi, R. B. Staszewski, H. A. R. Homulle, B. Patra, J. P. G. van Dijk, R. M. Incandela, L. Song, and B. Valizadehpasha, "15.5 Cryo-CMOS circuits and systems for scalable quantum computing," in *2017 IEEE ISSCC*, 2017, pp. 264–265.

[91] R. K. Gupta, S. Mitra, and P. Gupta, "Variability expeditions: a retrospective," *IEEE Design Test*, vol. 36, no. 1, pp. 65–67, 2019.

[92] A. Grill, E. Bury, J. Michl, S. Tyaginov, D. Linten, T. Grasser, B. Parvais, B. Kaczer, M. Waltl, and I. Radu, "Reliability and variability of advanced CMOS devices at cryogenic temperatures," in *2020 IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020.

[93] M. E. P. V. Zurita, L. L. Guevel, G. Billiot, A. Morel, X. Jehl, A. G. M. Jansen, and G. Pillonnet, "Cryogenic current steering DAC with mitigated variability," *IEEE J. Solid-State Circuits*, vol. 3, pp. 254–257, Jul. 2020.

[94] M. Onabajo, D. Gómez, E. Aldrete-Vidrio, J. Altet, D. Mateo, and J. Silva-Martinez, "Survey of robustness enhancement techniques for wireless systems-on-a-chip and study of temperature as observable for process variations," *J. Electron. Testing*, vol. 27, no. 3, pp. 225–240, Feb. 2011.

[95] G. S. May and C. J. Spanos, *Fundamentals of semiconductor manufacturing and process control*. John Wiley & Sons, 2006.

[96] G. Avenier, P. Chevalier, A. Montagne, L. Parmigiani, L. Berthier, O. Renault, E. Oghdayan, M. Mathieu, Y. Campidelli, D. Dutartre, and A. Chantre, "HBT device robustness against process variations in millimeter-wave BiCMOS technology," in *2012 IEEE Bipolar/BiCMOS Circuits and Technol. Meeting (BCTM)*, Sep. 2012.

[97] B. Anderson and R. Anderson, *Fundamentals of semiconductor devices*. McGraw-Hill, Inc., 2004.

[98] D. Klaassen, "A unified mobility model for device simulation—I. model equations and concentration dependence," *Solid-State Electron.*, vol. 35, no. 7, pp. 953–959, Jul. 1992.

[99] A. Schenk, P. Altermatt, and B. Schmithusen, "Physical model of incomplete ionization for silicon device simulation," in *2006 Int. Conf. Simul. Semicond. Processes and Devices (SISPAD)*, Sep. 2006.

[100] D. Roulston, N. Arora, and S. Chamberlain, "Modeling and measurement of minority-carrier lifetime versus doping in diffused layers of n^+-p silicon diodes," *IEEE Trans. Electron Devices*, vol. 29, no. 2, pp. 284–291, Feb. 1982.

[101] S. Jain and D. Roulston, "A simple expression for band gap narrowing (BGN) in heavily doped Si, Ge, GaAs and Ge_xSi_{1-x} strained layers," *Solid-State Electron.*, vol. 34, no. 5, pp. 453–465, May 1991.

[102] A. Schenk, "Finite-temperature full random-phase approximation model of band gap narrowing for silicon device simulation," *J. Appl. Phys.*, vol. 84, no. 7, pp. 3684–3695, Oct. 1998.

[103] G. L. Bir and G. E. Pikus, *Symmetry and strain-induced effects in semiconductors*. Wiley New York, 1974, vol. 484.

[104] H. Rucker, B. Heinemann, R. Barth, and M. Lisker, "Impact of back end stress on SiGe bipolar transistors," in *2011 IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2011.

[105] E. Canderle, P. Chevalier, G. Avenier, N. Derrier, D. Celi, and C. Gaquiere, "Impact of BEOL stress on BiCMOS9MW HBTs," in *2013 IEEE Bipolar/BiCMOS Circuits and Technol. Meeting (BCTM)*, Sep. 2013.

[106] I. D. Wolf, "Micro-Raman spectroscopy to study local mechanical stress in silicon integrated circuits," *Semicond. Sci. and Technol.*, vol. 11, no. 2, pp. 139–154, Feb. 1996.

[107] R. Shankar, *Principles of quantum mechanics*. Springer Science & Business Media, 2012.

[108] J. J. Wortman, J. R. Hauser, and R. M. Burger, "Effect of mechanical stress on p-n junction device characteristics," *J. Appl. Phys.*, vol. 35, no. 7, pp. 2122–2131, Jul. 1964.

[109] K. Dombrowski, A. Fischer, B. Dietrich, I. D. Wolf, H. Bender, S. Pochet, V. Simons, R. Rooyackers, G. Badenes, C. Stuer, and J. V. Landuyt, "Determination of stress in shallow trench isolation for deep submicron MOS devices by UV Raman spectroscopy," in *1999 IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 1999.

[110] M. Ieong, P. Solomon, S. Laux, H.-S. Wong, and D. Chidambarrao, "Comparison of raised and Schottky source/drain MOSFETs using a novel tunneling contact model," in *1998 IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 1999.

[111] P. Candra, V. Jain, P. Cheng, J. Pekarik, R. Camillo-Castillo, P. Gray, T. Kessler, J. Gambino, J. Dunn, and D. Harame, "A 130nm SiGe BiCMOS technology for mm-wave applications featuring HBT with f_T/f_{MAX} of 260/320 GHz," in *Proc. 2013 IEEE Radio Freq. Integr. Circuits Symp.*, 2013, pp. 381–384.

[112] E. Charbon, F. Sebastiano, A. Vladimirescu, H. Homulle, S. Visser, L. Song, and R. M. Incandela, "Cryo-CMOS for quantum computing," in *Proc. 2016 IEEE Int. Electron Devices Meeting*, 2016, pp. 13.5.1–13.5.4.

[113] D. J. Reilly, "Engineering the quantum-classical interface of solid-state qubits," *npj Quantum Information*, vol. 1, p. 15 011, 2015.

[114] J. D. Cressler and G. Niu, *Silicon-germanium heterojunction bipolar transistors*. Artech house, 2003.

[115] M Schroter and A Chakravorty, "Compact hierarchical modeling of bipolar transistors with HICUM," *World Scientific, Singapore*, 2010.

[116] S. Weinreb, J. C. Bardin, and H. Mani, "Design of cryogenic SiGe low-noise amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 55, no. 11, pp. 2306–2312, 2007.

[117] J. C. Bardin and S. Weinreb, "Experimental cryogenic modeling and noise of SiGe HBTs," in *2008 IEEE Int. Microw. Symp. Dig.*, 2008, pp. 459–462.

[118] B. O. Woods, H. A. Mantooth, and J. D. Cressler, "SiGe HBT compact modeling for extreme temperatures," in *2007 Int. Semicond. Devices Res. Symp.*, 2007, pp. 1–2.

[119] Z. Xu, G. Niu, L. Luo, and J. Cressler, "A physics-based trap-assisted tunneling current model for cryogenic temperature compact modeling of SiGe HBTs," ECS, 2010.

[120] L. Luo, Z. Xu, G. Niu, P. S. Chakraborty, P. Cheng, D. Thomas, K. Moen, J. D. Cressler, M. Mudholkar, and H. A. Mantooth, "Wide temperature range compact modeling of SiGe HBTs for space applications," in *Proc. 2011 IEEE Souteastern Symp. Syst. Theory*, 2011, pp. 110–113.

[121] M. Goryachev, S. Galliou, and P. Abbe, "Cryogenic transistor measurement and modeling for engineering applications," *Cryogenics*, vol. 50, no. 6, pp. 381 –389, 2010.

[122] M. Schroter, S. Lehmann, H. Jiang, and S. Komarow, "HICUM/level0 - a simplified compact bipolar transistor model," in *Proc. 2002 IEEE Bipolar Circuits and Technol. Meeting*, 2002, pp. 112–115.

[123] Z. Huszka, D. Cli, and E. Seebacher, "A novel low-bias charge concept for HBT/BJT models including heterobandgap and temperature effects-part ii: implementation, parameter extraction and verification," *IEEE Trans. Electron Devices*, vol. 58, no. 2, pp. 357–363, 2011.

[124] M. Müller and M. Krattenmacher, "Open license parameter extraction tool - overview and demo for SiGe HBTs," in *2019 HICUM Workshop*, 2019.

[125] A. Hastings, *The art of analog layout*. Prentice Hall, New Jersey, 2001.

[126] Y. Taur and T. H. Ning, *Fundamentals of modern VLSI devices*. Cambridge university press, 2013.

[127] Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor*. Oxford Univ. Press, 2011.

[128] T. Grotjohn and B. Hoefflinger, "A parametric short-channel MOS transistor model for subthreshold and strong inversion current," *IEEE J. Solid-State Circuits*, vol. 19, no. 1, pp. 100–112, 1984.

[129] W Fikry, G Ghibaudo, and M Dutoit, "Temperature dependence of drain-induced barrier lowering in deep submicrometre mosfets," *Electronics Letters*, vol. 30, no. 11, pp. 911–912, 1994.

[130] Y. Liu, L. Lang, Y. Chang, Y. Shan, X. Chen, and Y. Dong, "Cryogenic characteristics of multinanoscales field-effect transistors," *IEEE Trans. Electron Devices*, pp. 1–8, 2020.

[131] V. Jain, B. Zetterlund, P. Cheng, R. A. Camillo-Castillo, J. J. Pekarik, J. W. Adkisson, Q. Liu, P. B. Gray, V. Kaushal, T. Kessler, and D. Harame, "Study of mutual and self-thermal resistance in 90nm SiGe HBTs," in *2013 IEEE Bipolar/BiCMOS Circuits and Technol. Meeting (BCTM)*, 2013, pp. 17–20.

[132] F. J. De la Hidalga, M. J. Deen, and E. A. Gutierrez, "Theoretical and experimental characterization of self-heating in silicon integrated devices operating at low temperatures," *IEEE Trans. Electron Devices*, vol. 47, no. 5, pp. 1098–1106, 2000.

[133] J. C. S. Woo, J. D. Plummer, and J. M. C. Stork, "Non-ideal base current in bipolar transistors at low temperatures," *IEEE Trans. Electron Devices*, vol. 34, no. 1, pp. 130–138, 1987.

[134] H. Tuinhout, A. Bretveld, and W. Peters, "Measuring the span of stress asymmetries on high-precision matched devices," in *2004 IEEE Int. Conf. Microelectronic Test Structures*, Mar. 2004.

[135] N. Wils, H. P. Tuinhout, and M. Meijer, "Characterization of STI edge effects on CMOS variability," *IEEE Trans. Semicond. Manuf.*, vol. 22, no. 1, pp. 59–65, 2009.

[136] Y.-M. Sheu, K.-W. Su, S. Tian, S.-J. Yang, C.-C. Wang, M.-J. Chen, and S. Liu, "Modeling the well-edge proximity effect in highly scaled MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2792–2798, Nov. 2006.

[137] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous univariate distributions, volume 1*. John wiley & sons, 1995, vol. 289.

VITA

Hanbin Ying (nickname Victor) was born in Fuzhou, Fujian, China in 1994. After finishing the high school in China, he came to the United States for college in 2012. As an undergraduate researcher, he worked in Professor Ying Zhang's lab in 2014 and subsequently joined Professor John Cressler's lab in 2015. He received a B.S. degree in Electrical and Computer Engineering in 2016 with a full GPA, and a Master degree in Electrical and Computer Engineering in 2019, both from Georgia Institute of Technology, USA. His Ph.D. research focuses on using silicon germanium heterojunction bipolar transistors for emerging cryogenic applications such as quantum computing and superconducting nanowire single-photon detector readout. The research spans a broad area, including novel device physics at cryogenic temperature, construction of compact models for cryogenic SiGe HBTs, and circuit designs of SiGe HBTs at deep-cryogenic temperatures. He has authored and co-authored more than 20 publications during his Ph.D. research.

Hanbin received ECE Outstanding Junior Scholar Award at Georgia Tech in 2015, ECE Senior Scholar Award in 2016, Faculty Honors every year from 2012 to 2016, Georgia Tech Leadership Fellows in 2019, and the Best Student Paper Award (author) in *2019 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium*. Hanbin was the lab manager for the SiGe Devices and Circuit Team during 2018-2019.

Outside of research, Hanbin's interest includes reading, classical music, workout, and personal finance.